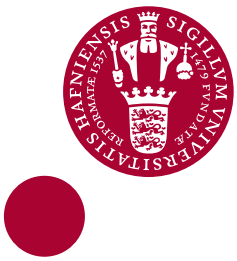

Comprehensive profiling of the fission yeast transcription start site activity during stress and media response

Thodberg, Malte; Thieffry, Axel; Bornholdt, Jette; Boyd, Mette; Holmberg, Christian; Azad, Ajuna; Workman, Christopher T; Chen, Yun; Ekwall, Karl; Nielsen, Olaf; Sandelin, Albin

Published in:
Nucleic Acids Research

DOI:
10.1093/nar/gky1227

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

Published online 19 December 2018

Nucleic Acids Research, 2019, Vol. 47, No. 4 ***1671–1691***
doi: 10.1093/nar/gky1227

Comprehensive profiling of the fission yeast transcription start site activity during stress and media response

Malte Thodberg[1], Axel Thieffry[1], Jette Bornholdt[1], Mette Boyd[1], Christian Holmberg[2], Ajuna Azad[1], Christopher T. Workman[3], Yun Chen[1], Karl Ekwall[4], Olaf Nielsen[2,*] and Albin Sandelin[1,*]

[1]Department of Biology and Biotech Research and Innovation Centre, The Bioinformatics Centre, University of Copenhagen, DK2100 Copenhagen N, Denmark, [2]Department of Biology, Cell cycle and genome stability Group, University of Copenhagen, DK2100 Copenhagen N, Denmark, [3]Department of Biotechnology and Biomedicine, Technical University of Denmark, DK2800 Kongens Lyngby, Denmark and [4]Department of Biosciences and Nutrition, Karolinska Institute, SE14183 Huddinge, Sweden

Received August 09, 2018; Revised November 09, 2018; Editorial Decision November 24, 2018; Accepted November 26, 2018

ABSTRACT

Fission yeast, *Schizosaccharomyces pombe,* is an attractive model organism for transcriptional and chromatin biology research. Such research is contingent on accurate annotation of transcription start sites (TSSs). However, comprehensive genome-wide maps of TSSs and their usage across commonly applied laboratory conditions and treatments for *S. pombe* are lacking. To this end, we profiled TSS activity genome-wide in *S. pombe* cultures exposed to heat shock, nitrogen starvation, hydrogen peroxide and two commonly applied media, YES and EMM2, using Cap Analysis of Gene Expression (CAGE). CAGE-based annotation of TSSs is substantially more accurate than existing PomBase annotation; on average, CAGE TSSs fall 50–75 bp downstream of PomBase TSSs and co-localize with nucleosome boundaries. In contrast to higher eukaryotes, dispersed TSS distributions are not common in *S. pombe*. Our data recapitulate known *S. pombe* stress expression response patterns and identify stress- and media-responsive alternative TSSs. Notably, alteration of growth medium induces changes of similar magnitude as some stressors. We show a link between nucleosome occupancy and genetic variation, and that the proximal promoter region is genetically diverse between *S. pombe* strains. Our detailed TSS map constitutes a central resource for *S. pombe* gene regulation research.

INTRODUCTION

Yeast cells have been central models for understanding eukaryotic gene regulation. Historically, baker's yeast (*Saccharomyces cerevisiae*) has been the unicellular model of choice, but the remotely related fission yeast *Schizosaccharomyces pombe* has in many ways turned out to be a more relevant model for mammalian cells (1). The architecture of *S. pombe* and human chromosomes are similar, featuring large repetitive centromeres and regions of RNAi- dependent heterochromatin (2). *Schizosaccharomyces pombe* also utilizes histone modifications and chromatin remodeling enzymes similar to multicellular eukaryotes (2). On the other hand, both yeast types have highly related signal transduction pathways responding to environmental stress (3).

A central challenge for cells is to react to changing environments through the regulated activation of gene transcription. The most commonly studied gene regulation responses in yeast cells are to environmental stress through chemicals (e.g. hydrogen peroxide, sorbitol, cadmium), physical conditions (e.g. heat), or changes in available nutritional substances (nitrogen starvation, change of growth media, etc.). Previous work on stress response in *S. pombe* and *S. cerevisiae* has focused on the distinction between a general environmental stress response (Core Environmental Stress Response, CESR) versus specific response to individual types of stress (Specific Environmental Stress Response, SESR). CESR is comprised of metabolic genes related to carbohydrate metabolism and genes involved in protein stability such as anti-oxidants, proteases and heat shock proteins, while SESR is comprised of genes with more specific functions related to the given type of stress (4–7).

*To whom correspondence should be addressed. Tel: +45 3532 1281; Fax: +45 3532 1281; Email: albin@binf.ku.dk
Correspondence may also be addressed to Olaf Nielsen. Tel: +45 26 41 06 66; Fax: +45 3532 1281; Email: onigen@bio.ku.dk

In *S. pombe*, the timing and maintenance of the CESR is controlled by a conserved signal transduction pathway that ultimately activates the Sty1 protein kinase (8), a member of the stress-activated MAP kinase family (SAPK family), which also includes HOG1 from *S. cerevisiae* and human *MAPK14* and cJun-N terminal kinase (*MAPK8*). Sty1, in turn, activates key transcription factors such as Atf1, Pap1 and Hsf1 (8–10).

Stress-specific gene regulation has been studied by comparing genome-wide changes in expression profiles in wild-type cells and mutants lacking key signaling components or transcription factors (4,11). While some types of stress (for example heat shock) induce a quick and transient response, hydrogen peroxide and alkylating agents induce a more lengthy response (4). Furthermore, both stress exposure and environmental cues, such as nutritional composition, modulate the growth rate and size at which fission yeast cells enter mitosis (12). This size regulation occurs partly through the Sty1 SAPK and also via the TOR (target of rapamycin) pathway (13,14). Extreme nutritional stresses, in particular nitrogen starvation, induces cells to enter sexual development, a process which is intertwined with the CESR (15,16).

Accurate maps of TSSs and their activity have been instrumental in understanding gene regulation and core promoter activity, as well as the evolution of gene regulation in mammals, birds, insects and plants (e.g. (17–22)). Such data sets have also been highly beneficial in deciphering the relationship between chromatin and transcription initiation (e.g. (19,23–25)). Due to its importance as a model organism in chromatin biology and stress response research, it is surprising that no comprehensive maps of transcription start sites (TSSs) across different cellular states have been reported for *S. pombe*.

Cap Analysis of Gene Expression (CAGE), based on sequencing the first 20–30 bp of capped, full-length cDNAs (26) is arguably the most used technique for locating TSSs and their transcriptional activity genome-wide (27). CAGE tags can, when mapped to a reference genome, identify TSSs with single nucleotide resolution and quantify their level of transcription, as the number of CAGE tags mapping to a location is proportional to the concentration of the originating mRNA (Figure 1A). CAGE tags can thus be used for expression profiling, but on TSS rather than gene level. In this sense, it is complementary to RNA-Seq, which has the advantage that splicing can be assessed but, on the other hand, is not precise in locating TSSs. Previous studies have shown that RNA-Seq and CAGE have comparable accuracy in terms of expression profiling (28). Because CAGE is not contingent on existing gene models, it can both refine existing gene models and locate novel TSSs, both within and outside of known genes. In eukaryotes, alternative TSS usage is common (29): for instance, in human and mouse >50% of genes have two or more distinct TSSs, many of which are used in a tissue- or context-specific manner (22). Such alternative TSSs are interesting as they may confer additional, independent regulatory inputs for genes and/or change the protein-coding potential of the resulting mRNA, for instance by excluding exons coding for protein domains (30,31).

Since stress response is highly studied in *S. pombe*, having accurate and genome-wide TSS maps for stress states would be beneficial for understanding their gene regulation and associated processes. Here, we used CAGE to define a TSS atlas of unprecedented resolution and scope for *S. pombe*, across three environmental stressors (heat shock, nitrogen starvation and hydrogen peroxide stress) and two commonly used growth media, Edinburgh Minimal Medium (EMM2) and Yeast Extract Medium with supplements (YES). We show that this atlas substantially expands and refines current state-of-the-art *S. pombe* TSS annotation, allowing for more detailed interpretation of a range of processes, including nucleosome positioning, histone modification and transcription levels. Additionally, our CAGE-based annotation allows the analysis of stress-specific and shared stress response regulation and growth media adaptation, and enables refined analysis of *S. pombe* genetic variation data.

MATERIALS AND METHODS

Culture conditions and RNA preparation

Triplicate cultures of the wild-type strain h^{-S} (972) were grown at 30°C in either EMM2 (Edinburgh minimal medium) or YES (yeast extract with supplements) to a density of 5×10^6 cells/ml (32). For nitrogen starvation, cells were transferred by vacuum filtration from EMM2 to EMM2 without ammonium chloride and incubated for 16 h at 30°C. Heat shock was imposed by transferring YES cultures to 39°C for 15 min. Oxidative stress was inflicted by treating YES cultures with 0.5 mM H_2O_2 for 15 min at 30°C. Total RNA was extracted from 10^8 cells. In brief, cells were harvested by 5 min of centrifugation at 3000 rpm. Pellets were resuspended in TES (10 mM TrisHCl pH7.5, 10 mM ethylenediaminetetraacetic acid and 0.5% sodium dodecyl sulphate) and transferred to 65°C preheated acidic phenol-chloroform (Sigma P-1944). After 1 h of incubation at 65°C with mixing every 10 min, RNA was extracted with chloroform-isoamyl alcohol (Sigma C-0549), ethanol precipitated and re-suspended in water. All RIN-values were above 8.8. As CAGE requires a certain amount of input RNA concentration, RNA from nitrogen-starved cultures was additionally concentrated by vacuum centrifuge (reported RIN-scores are after concentration). See Supplementary Table S1 for an overview of libraries.

CAGE analysis and mapping

CAGE libraries were prepared from the 15 yeast cultures as in (26), using 5 μg total RNA as starting material. Libraries were run individually with the following four barcodes: #2(CTT), #3 (GAT), #4 (CACG) and #8 (ATC). Prior to sequencing, four CAGE libraries with different barcodes were pooled and applied to the same sequencing lane. Sequencing of the libraries was performed on a HiSeq2000 instrument from Illumina at the National High-Throughput DNA Sequencing Centre, University of Copenhagen. To compensate for the low complexity in 5′ ends of the CAGE libraries, 30% Phi-X spike-ins were added to each sequencing lane, as recommended by Illumina. CAGE reads were assigned to their respective originating sample according to

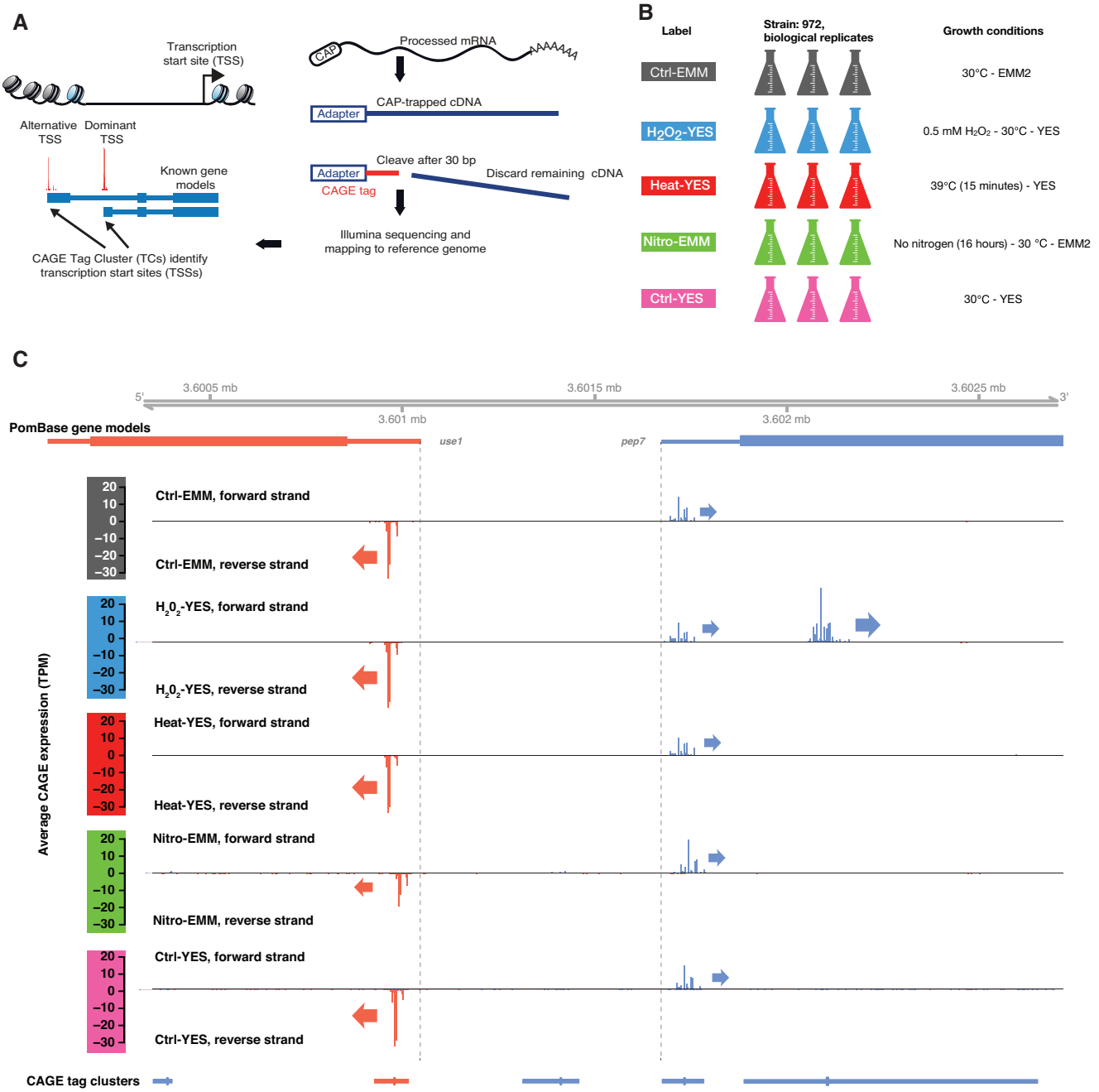

Figure 1. Overview of CAGE experiment. (**A**) Schematic overview of the CAGE protocol. Capped full-length cDNAs are isolated from total RNA through cap-trapping, followed by cleaving off the first 30 bp, which are sequenced and mapped back to the reference genome, identifying TSS locations and their relative usage. Such 30 bp reads are referred to as CAGE tags. Nearby CAGE tags on the same strand are grouped into TCs. (**B**) Experimental design. Five sets of biological *Schizosaccharomyces pombe* triplicates were prepared, covering three stressors (heat shock, hydrogen peroxide stress and nitrogen starvation) and two common growth media: Edinburgh Minimal Media (EMM2) and Yeast Extract (YES). Labels for samples/libraries are shown in the left column. Growth conditions are shown in the right column. (**C**) Genome-browser example of CAGE defined TSS landscape of the *pep7* gene locus. *X*-axis represents the genome. Panels represent different types of data mapped to the genome. Data on forward strand is indicated by blue color, reverse strand by red. The first and second tracks show PomBase gene models; dotted vertical lines show PomBase-defined TSSs. The third to seventh tracks show the average pooled CAGE signal from each experimental group (indicated by color code to the left) as bar plots where the Y-axis shows CAGE expression (TPM-normalized number of 5′-ends of CAGE tags at a given bp), where positive values and blue color indicate forward strand CAGE-tags, and negative values and red color indicate reverse strand CAGE-tags. Arrows show the direction of transcription for highly used TSSs. The last track shows the location of clusters of nearby CAGE tags, where thicker lines indicate the TSS peaks, i.e. the position with the highest CAGE signal. We identified ubiquitously expressed TSS for *use1* and *pep7* located just downstream of the PomBase gene model-defined TSSs, but also a novel *pep7* alternative TSS only active in H_2O_2-YES. Also, see Supplementary Figure S1B–E for additional genome browser examples.

identically matching barcodes. Using the FASTX Toolkit (v0.0.13), assigned reads were: (i) 5′-end trimmed to remove linker sequences (10+1 bp to account for the CAGE protocol G-bias at the first 5′ base), (ii) 3′-end trimmed to a length of 25 bp and (iii) filtered for a minimum sequencing quality (Phred score) of 30 in 50% of the bases. Trimmed reads were mapped using Bowtie (33) (version 0.12.7) with parameters –t –best –strata –v –k 10 –y –p 6 –phred33-quals –chunksmbs 512 –e 120 –q –un to ASM294v2.26. To obtain bp resolution tracks, the number CAGE tag 5′-ends were counted for each genomic position. Such CAGE-supported bp are often referred to as CTSSs in previous works (22), but will here be referred to as bp-resolution TSSs.

These coordinates were offset by 1 bp to account for the G-bias trimming. Only chromosomes I, II and III were used for analysis. Mapping statistics are available in Supplementary Table S1.

Quantification and annotation of TSSs and genes

For each library, bp-resolution TSS counts were normalized into tag-per-million (TPM) values and the sum of TPM-values across all libraries were calculated for each base pair. Nearby bp resolution TSSs within 20 bp of each other were merged into Tag Clusters (TCs). Only bp resolution TSSs passing a TPM-threshold (0.04516233) were considered; this threshold was chosen to maximize the total number of TCs across the genome (as implemented in the CAGEfightR R-package version 0.3, https://bioconductor.org/packages/CAGEfightR/). Expression was quantified as the number of tags in each TC for each sample. TCs having > 1 TPM in at least three libraries were retained. In the main text, we refer to TCs as 'CAGE-defined TSS' for simplicity, but use TC nomenclature in Methods to be comparable to older literature. Transcript- and gene-model annotations (ASM294v2.26) were downloaded from PomBase as a GTF-file and imported as a TxDb object (34). Regions were extracted using the *transcriptsBy*-family of functions and genes were extracted using the *genes*-function. TCs were annotated based on their overlap with annotated transcripts using the hierarchical models shown in Figure 2A: promoters were defined as the ±100 bp region around a PomBase TSS, while Proximal was defined as −1000 to −100 from annotated TSS. 5′ UTRs, 3′ UTRs and CDS regions were defined as in PomBase. Exonic refers to non-protein-coding exons, including lncRNAs. Antisense was defined as intragenic but on the opposite strand. To assess gene-level expression, TCs were aggregated by summing all counts within or −1000 upstream of annotated genes. In case a TC overlapped more than one gene, the gene with the nearest annotated TSS was chosen.

To make TSSs obtained from PomBase comparable to TCs in terms of expression, the same approach was used: CAGE signal at bp resolution TSSs were quantified around PomBase TSSs (first bp of the 5′-UTR) ±100 bp and were then TPM-filtered as above. Bp resolution TSSs, TCs and transcript models were visualized using CAGEfightR and the Gviz R-package (35). TCs and bp-resolution tracks for each condition are deposited in the GEO database (GSE110976).

Biological sequence analysis

The *getSeq*-function from the BSgenome package was used to extract sequences from the reference genome. Di- and trinucleotide frequencies were counted using *vmatchPattern*-function and sequence logos were made using the ggseqlogo package (36). TATA-Box and INR-motifs were obtained from JASPAR (37) via the JASPAR2016 R-package (38). The *motifScanScores*-function from the seqPattern R-package (http://bioconductor.org/packages/seqPattern/) was used to scan for motif occurrences. CpG frequency (CpG dinucleotides per bp) was calculated in a −100/+100 window around TSS peaks.

Average CAGE, MNase-seq, H3K4me3 ChIP-seq, PRO-Cap, PRO-Seq and NET-seq signal calculations

MNase-seq and H3K4me3 ChIP-Seq data was obtain from (39,40) (accession numbers GSM1374060, GSE49575). Only wild-type strains were used in both cases. PRO-Cap and PRO-Seq signals were obtained from (41) (accession number GSE76142). NET-seq data were obtained from (42) (BAM-files were generously supplied by the authors). The GenomicAlignments package (34) was used to import and calculate the average signal.

In all cases, data were imported with rtracklayer and coverage calculated with the *coverage*-function from IRanges (34). CAGE, PRO-Cap, PRO-Seq and NET-Seq profiles were plotted after removing the 1% most highly expressed sites, to remove the influence of a few extreme outliers.

For Figure 3A and B, only CAGE-defined TSSs at annotated promoters were used. For Figure 3C and Supplementary Figure S3C and D, TSSs were selected based on the following criteria: (i) the TSSs must have an upstream CAGE TSS within 250 bp on the opposite strand and (ii) the TSSs must not be within 500 bp of an annotated TSS, on either strand. TSSs were grouped by the amount of NET-Seq sense signal in the nucleosome-depleted region (NDR), defined as −250:−50 bp relative to the CAGE-defined TSS peak.

Comparison with TSS atlases from Li *et al.* and Eser *et al.*

Eser *et al.* (43) TSSs were obtained from the transcriptional units from the supplementary material of the paper by selecting the first 5′-end position as the TSSs. Li *et al.* (44) bp resolution TSSs and TCs were graciously provided by the authors. We retained Li *et al.* TCs with >1 TPM and recalculated TC peaks similar to above to obtain TSSs.

Comparison with mammalian TSSs

We obtained bp resolution TSSs from five control replicates for CAGE assays on mouse lung from (45). As these libraries were prepared in the same lab as the *S. pombe* libraries presented here and analyzed using the same pipeline, they are highly comparable. TC widths were calculated as the distance between the two positions in the TC corresponding to the 10th and 90th percentile of the pooled TPM across all libraries.

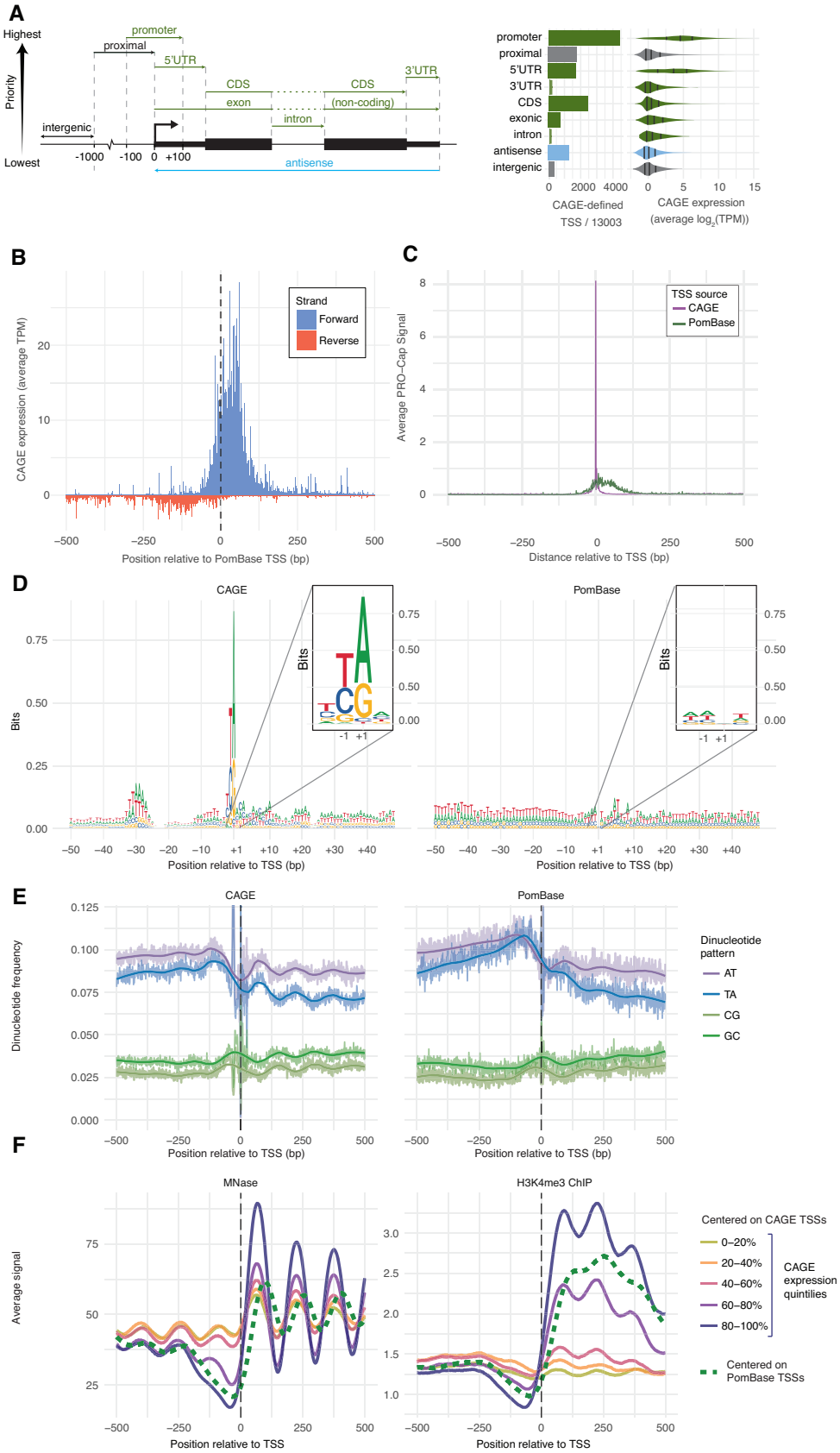

Figure 2. Properties of CAGE-defined TSS and comparison to PomBase annotation. (**A**) Location of CAGE-defined TSS versus gene annotation. Left panel: CAGE-defined TSSs were grouped into one of nine categories based on overlap with PomBase reference annotation (see schematic on the left:

Differential expression analysis and GO-term enrichment

All analysis was carried out using the edgeR (46) and limma (47) R-packages. The same analysis was run for TSSs and for genes. Initial normalization using the TMM-method from edgeR revealed a consistent shift in the overall distribution of expression values in the Nitro-EMM libraries. This could either reflect a large global shift in transcriptional output or an artifact of the additional RNA-purification steps in these samples (see above). Subsequent analysis using only TMM-normalization revealed a large number of upregulated compared to downregulated TSSs as well as a global shift in fold changes for highly expressed genes. For the final analysis we therefore chose the conservative approach of also normalizing expression using the cyclicloess-method prior to voom-transformation to model the mean variance relationship inherent to count data. Heat maps are based on the dual-normalized expression values. Multidimensional scaling (MDS) plots were made using the *plotMDS*-function from the limma package using pairwise Euclidian distances between the top 1000 genes/TSSs (top = 1000) and otherwise standard settings.

We modeled expression using ∼treatment+medium with Ctrl-EMM libraries as reference levels. We used robust estimation at the empirical Bayes stage and used the treat method to test against a minimum required $\log_2$ fold change of 0.5 (corresponding to ∼41% higher or lower expression between states). Resulting *P*-values were corrected for multiple testing using Benjamini–Hochberg. Supplementary Figures S4 and S5E–J show diagnostic plots for the differential expression analysis on TSS and gene level.

Housekeeping candidates were identified using an F-test across all coefficients, with resulting *P*-values corrected for multiple testing using Benjamini–Hochberg. Biological Coefficient of Variations (BCVs) were calculated using the *estimateDisp*-function with robust = TRUE from the edgeR package. Genes with low BCV were identified by splitting TSSs/genes into 25 bins based on average expression, and selecting the 10% TSSs/genes with the lowest BCV.

Differential transcript usage was assessed using the diffSplice-approach, with the Simes-method to aggregate results at the gene level. Resulting *P*-values were corrected for multiple testing using the Benjamini–Hochberg method.

GO-terms were downloaded from the PomBase, and the kegga-function was used to test for enrichment, with mean expression used as the covariate. Resulting *P*-values were corrected for multiple testing using Benjamini–Hochberg.

An alternative parameterization of the differential expression analysis based on pairwise comparisons was additionally carried out: The same limma-voom approach was used, but instead a design of ∼0+group was used and all pairwise comparisons extracted using contrasts. This allows for inspection of comparisons not considered in the main analysis, e.g. Nitro-EMM versus Heat-YES. Results are available in Supplementary Table S4.

Comparison with previously defined gene sets (CESR, SESR and nitrogen starvation) and alternative 5′ UTR TSSs

CESR and SESR (Peroxide and Heat shock) genes were downloaded from (4) (http://bahlerweb.cs.ucl.ac.uk/projects/stress/) and Nitrogen Starvation genes were downloaded from (15) (http://bahlerweb.cs.ucl.ac.uk/projects/sexualdifferentiation/meiosis/). Genes were matched to the CAGE data based on gene IDs. $\log_2$ fold changes with 95% confidence intervals estimated with limma-voom were plotted for each gene.

Genes utilizing alternative 5′-UTR usage was obtained from (48). The 24/28 genes could be matched by gene IDs. As the original data set did not provide coordinates for novel TSSs, we compared our gene-level differential transcript usage results from limma-voom with diffSplice to investigate agreements.

Motif enrichment analysis in promoter regions

Promoter region DNA sequences for TSSs were extracted as sequences −300 to +100 of TSS peaks (Supplementary Figure S1A) via BSgenome and *getSeq* as described above. We used Homer (49) in *de novo* pattern discovery mode to detect enriched motifs in each differentially expressed set compared to a background set consisting of TSSs not part of any differentially expressed set (default setting except size = given, S = 9, *P* = 40 and mset = yeast). Because Homer works without prior knowledge of DNA-binding motifs, identified motifs may be novel patterns or correspond to binding preferences of known transcription factors. To link identified motifs with known motifs, we compared them to Homer's database of *S. cerevisiae* motifs. The top 5 enriched *de novo* motifs with a foreground set frequency of >0.9%

←

categories are ordered by priority: in case a CAGE-defined TSS overlapped two or more categories, it was assigned to the one with the highest priority). Colors indicate whether regions overlap genes on same strand, antisense strand or are intergenic. Middle panel shows a bar plot counting the number of CAGE-defined TSS assigned to each category (from a total of 13 003 expressed TSSs). Right panel shows the distribution of CAGE expression within each category (average TPM across all replicates and conditions) as violin plots, colored as above. Marks within distributions indicate first quantile, median and third quantile. (**B**) Comparison of CAGE and PomBase TSS locations. *X*-axis show distance relative to annotated PomBase TSSs in bp. *Y*-axis shows average CAGE signal as TPM across all libraries: CAGE tags on the reverse strand in relation to the annotated TSS are shown as negative values (red), CAGE tags on forward strand are shown as positive values (blue). Dotted line indicates PomBase TSS locations. (**C**) Comparison of PRO-Cap signal at PomBase or CAGE-defined TSSs. *X*-axis shows distance relative to TSSs defined by CAGE or PomBase. *Y*-axis shows the average PRO-Cap signal around CAGE-defined TSSs (purple) and annotated PomBase TSSs (green). (**D**) Comparison of CAGE and PomBase promoter patterns. Each panel shows a sequence logo of the ±50 bp genomic region around TSS peaks defined by CAGE (left panel) or PomBase annotation (right panel). *Y*-axis shows information content in bits. Colors indicate DNA-bases. Insets show a zoom-in of the ±2 region in each panel. (**E**) Comparison of CAGE and PomBase promoter di-nucleotide frequencies. *Y*-axis shows the frequency of di-nucleotides (as indicated by color) around TSSs defined by CAGE or PomBase; thick lines show loess regression trend lines. *X*-axes show distance relative to CAGE-defined TSSs (right panel) or PomBase-defined TSSs (right panel) in bp. Frequencies > 0.125 are cut. (**F**) Comparison of MNAse-Seq and H3K4me3 ChIP-Seq signals anchored at CAGE or PomBase TSS. *X*-axis indicates the distance from TSSs defined by CAGE or PomBase. *Y*-axis shows the average MNase-Seq or H3K4me3 ChIP-Seq signal, where larger values indicate highly positioned nucleosomes (left) or modified histones (right). Colors indicate whether signals are anchored on CAGE-defined TSSs (solid lines, stratified by expression quintiles) or PomBase-defined TSSs (dashed).

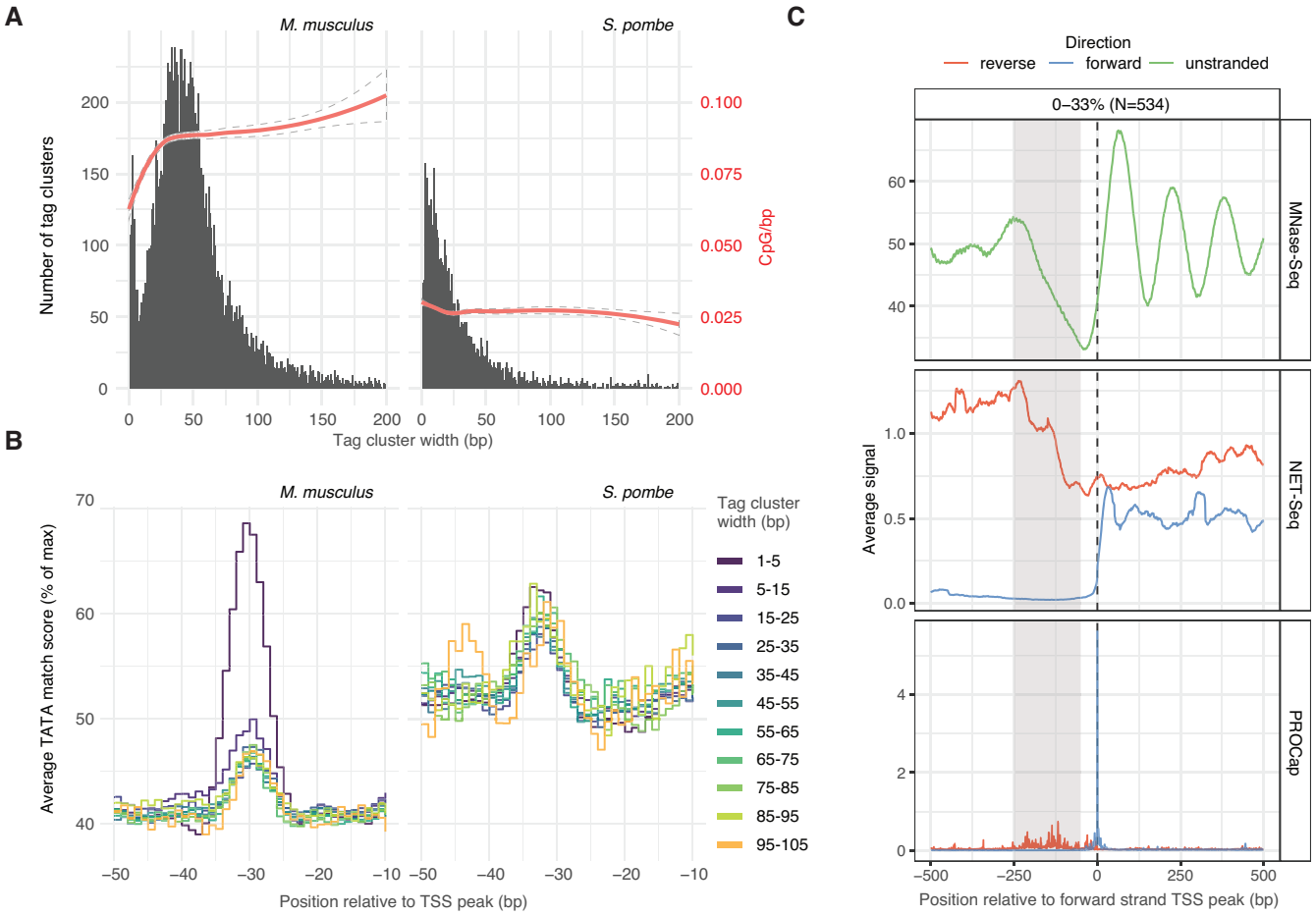

Figure 3. Analysis of TSS distribution shape and bidirectional transcription in *Schizosaccharomyces pombe.* (**A**) Relation between CAGE TC width and CpG content. *X*-axes show CAGE TC width as defined in Supplementary Figure S1A. Left *Y*-axes and black histogram show the distribution of widths in five pooled *Mus musculus* lung CAGE libraries (left panel) and this study (*S. pombe,* right panel). Red trend-lines and right *Y*-axes show the average number of CpG dinucleotides per bp (200 bp centered on TSS peaks). The 95% confidence intervals are shown as dotted lines. Only CAGE-defined TSSs corresponding to annotated TSSs were analyzed. (**B**) Relation between TATA-box occurrence and TC width. *X*-axes show the −50 to −10 region relative to TSS peaks. *Y*-axes show the average TATA-box motif match score (0–100% of maximal score). TSSs are stratified by their TC width, indicated by line color. CAGE-defined TSSs from *M. musculus* and *S. pombe* are shown in left and right panel, respectively. Only CAGE-defined TSSs corresponding to annotated TSSs are analyzed. (**C**) Analysis of bidirectional transcription initiation. *X*-axes show distance relative to *S. pombe* CAGE-defined TSSs in bp. Each panel shows a different assay (MNase-Seq, NET-Seq or PRO-Cap). *Y*-axis shows average signal, in forward (blue), reverse (red) direction or unstranded (green) relative to the TSS. TSSs analyzed were selected to have bidirectional transcription based on CAGE, and subsequently filtered to only retain those with little sense NET-Seq transcription in the NDR-region (highlighted in gray). Supplementary Figure S3C and D shows TSSs with higher NET-Seq signal in the NDR region.

were plotted using ggseqlogo. *S. cerevisiae* motifs was also used to perform known motif enrichment for the same sets (Dataset S1).

Single nucleotide variation analysis

Genetic variation data was downloaded from Ensembl-Fungi (50). Only Single Nucleotide Variations (SNVs) ('TSA = SNV;Jeffares_SNPs') were used for analysis. IRanges (34) were used to calculate the average number of SNVs per basepair in promoter regions (TSS peaks −300:+100 base pairs). GC-content and TATA-box positions were calculated as above. Background SNV frequencies were calculated as the total number of SNVs overlapping CDS regions divided by the total size in basepairs of all CDS regions. Filtering TSSs to only include TSSs at annotated promoters more than 500 bp distant from other such TSSs did not affect the results (data not shown).

The number of SNVs within promoter regions as a function of differential expression status and annotation category was modeled using negative binomial (NB) regression (due to overdispersion compared to a Poisson distribution) using the *glm.nb*-function from the MASS R-package (51). Exponentiated effect estimates and confidence intervals were extracted using the *tidy*-function from the broom R-package (https://CRAN.R-project.org/package=broom). Including average TSS expression in the model did not have any major effects (data not shown).

Protein domains in genomic space

Pfam protein domains (52) were obtained from PomBase. Amino acid positions were multiplied by three and mapped

back to genome using the *mapFromTranscripts*-function from GenomicFeatures (34). Domain disruptive TSSs were defined as TSSs that either: (i) overlapped a domain or (ii) were downstream of a domain within the same gene.

Northern blot analysis

Standard northern analysis was performed on the RNA samples used for CAGE analysis. The hybridization probe was obtained by labeling a 270-bp PCR fragment from exon 5 of *cds1*, generated with the primers Cds1-F: ttactgcgtctattccttttg and Cds1-R: cgaagaattgagctgttcg.

Western blot analysis

Protein extracts from a Cds1-HA tagged strain and a wild-type control treated as above (Ctrl –EMM, Nitro-EMM, Heat-YES and Ctrl-YES for both strains) were made by the TCA precipitation method. Extracts were fractionated by sodium dodecyl sulphate-polyacrylamide gel electrophoresis. After semi-dry transfer to a nitrocellulose membrane, Cds1 was detected using primary anti-HA (12CA5) monoclonal antibodies and HRP-coupled secondary anti-mouse antibodies.

RESULTS

A TSS atlas for *S. pombe*

We prepared three biological replicates of *S. pombe* cultures growing under five different conditions: cells growing exponentially in EMM2 medium (designated as 'Ctrl-EMM'), cells growing exponentially in YES ('Ctrl-YES'), heat-shocked cells growing in YES ('Heat-YES'), nitrogen starved cells growing in EMM2 ('Nitro-EMM') and hydrogen peroxide-stressed cells growing in YES ('H_2O_2-YES') (Figure 1B). After RNA purification, we constructed three CAGE libraries per condition, corresponding to each biological replicate (the CAGE method is outlined in Figure 1A). The average yield was 16.9 million uniquely mapping tags to the ASM294v2.26 assembly (Supplementary Table S1).

Mapped CAGE tags closely located to each other on the same strand were grouped into TCs, similarly to (53). The expression of TCs across libraries was assessed as the number of CAGE tags within each cluster, normalized by sequencing depth in the respective library as TPM. We retained only TCs that had at least 1 TPM in three or more libraries. After cutoffs, we detected 13 003 TCs. For simplicity, we will refer to these TCs as 'CAGE-defined TSSs'. For each of these 13 003 CAGE-defined TSSs, we identified the 'TSS peak' as the single bp position with the highest TPM value across all libraries (Supplementary Figure S1A).

In order to interpret the activity of CAGE-defined TSSs, it is useful to annotate them in reference to known transcripts and gene models. As an example, Figure 1C shows the annotated TSSs of the *use1* and *pep7* genes, which are positioned bidirectionally ∼1 kb apart. Both annotated TSSs were detected by CAGE in all conditions, albeit consistently slightly downstream of the annotated TSSs (see below for a systematical investigation of this phenomenon). However, an unannotated TSS for *pep7* was detected 106 bp downstream of the PomBase-annotated TSSs. Unlike the annotated TSS, this TSS was only strongly expressed in H_2O_2-YES libraries. It is thus an H_2O_2-YES-specific novel alternative TSS. We show genome-browser examples of key stress response genes in Supplementary Figure S1B–D, and a more complex locus with multiple TSSs and transcripts in Supplementary Figure S1E.

CAGE-defined TSSs refine PomBase TSS annotations

To systematically investigate the overlap with existing gene models, each CAGE-defined TSS was annotated into one of nine categories, depending on their overlap with transcript models as defined by PomBase (Figure 2A, left). CAGE-defined TSS overlapped PomBase-annotated promoter regions (±100 bp from the start of PomBase annotated TSSs) in 34% of cases (Figure 2A, right). Remaining CAGE-defined TSSs (8590 TSSs) commonly overlapped the proximal upstream region of genes (up to 1 kbp upstream of PomBase annotated TSSs, *14%*), 5′-UTRs (*13%*), and coding exons (CDS) (*18%*). Notably, CAGE-defined TSSs overlapping annotated promoters or 5′-UTRs were generally more highly expressed than CAGE-defined TSSs within other categories. This reflects observations in mouse and human (e.g. (45,53)) and indicates that novel TSSs on average are less expressed, although clear exceptions to this rule exist (e.g. Figure 1C).

We noted that while CAGE-defined TSSs often overlapped the regions around PomBase-annotated gene TSSs, the peak (the position having the most CAGE tags) of the corresponding CAGE TC was often located slightly downstream of PomBase TSSs (exemplified in Figure 1C). Plotting the average CAGE signal around PomBase-defined TSSs confirmed the existence of a small but consistent shift, where the majority of CAGE tags fell 50–75 bp downstream of annotated TSSs (Figure 2B and Supplementary Figure S2A). This difference could either be due to systematic overestimation of gene lengths in the PomBase annotation, or a systematic bias in our CAGE data. We reasoned that the most correct TSS set should better recapitulate known biology of gene regulation at DNA-sequence and chromatin levels.

First, we compared our data to an independent PRO-Cap dataset (41). Like CAGE, PRO-Cap is based on sequencing of capped 5′-ends but captures nascent RNA. CAGE TSSs coincided with extremely high and focused PRO-Cap peaks (Figure 2C) while most PRO-Cap signal was dispersed downstream of PomBase TSSs, consistent with the shift observed in the CAGE data (Figure 2B). Additionally, NET-Seq (42) and PRO-Seq (41) data, capturing 3′-ends of nascent RNAs within RNA Polymerase II (RNAPII), showed peaks +30 (NET-Seq) and ∼+55 (PRO-Seq) (Supplementary Figure S2B), likely corresponding to poised/stalled RNAPII that initiated transcription at the CAGE-defined TSSs.

Second, we investigated sequence content around TSSs. Previous experiments have shown that TATA and INR core promoter elements have a strong positional preference around TSS (∼−32/−25 for TATA and immediately at the TSS for INR) when present (54,55). A sequence logo (56) constructed by aligning the DNA sequence ±50 bp

around all peaks of CAGE-defined TSS (Figure 2D, left) showed enrichment of T/A-rich sequences at −32 to −25 and a strong pyrimidine-purine dinucleotide enrichment at −1/+1. These correspond to the TATA box and the central part of the INR element, and this pattern was highly similar to corresponding analysis in human and mouse (22,55). While this pattern was present across CAGE-defined TSSs in all annotation categories as defined above (Supplementary Figure S2C), it was not detected when constructing a sequence logo based on PomBase-defined TSSs (Figure 2D, right). Dinucleotide analysis in the same regions showed a 150-bp cyclical pattern with higher AT/TA frequency downstream of CAGE TSSs (Figure 2E, left). A similar, but weaker, pattern was present when centering on PomBase TSSs (Figure 2E, right). Nucleosome structure in *S. pombe* is highly DNA sequence dependent, determined by relative frequencies of CG and TA dinucleotides (57–59). We reasoned that if the 150-bp cyclical AT-rich patterns reflected nucleosomal placement, then nucleosome alignment should be stronger and better when positioned by CAGE data compared to PomBase. To test this, we analyzed nucleosomal positions by MNase-Seq (micrococcal nuclease digestion followed by sequencing) in Ctrl-EMM cells from (39). Indeed, MNase-Seq signal displayed the same cyclic pattern downstream of CAGE-based TSSs as predicted from dinucleotide frequencies, where the amplitude, but not phase, correlated with the CAGE expression strength and the position of the +1 nucleosome edge was adjacent to the CAGE TSS (Figure 2F, left and Supplementary Figure S2B, top right). H3K4me3 ChIP-seq (40) signals from corresponding cells showed similar patterns (Figure 2F, right, and Supplementary Figure S2B, bottom right).

At last, we compared our CAGE data with two previous TSS atlases for *S. pombe*: Li *et al.* (44) based on a single CAGE library and Eser *et al.* (43) based on two RNA-Seq libraries. Compared to both these resources, our CAGE-based TSS atlas contained thousands of novel TSSs (Supplementary Figure S2D). This likely reflected (i) our much larger sample size and higher library depth and (ii) that our atlas covered a large range of environmental conditions. The majority of Li *et al.* and Eser *et al.* TSSs were confirmed by our CAGE-based TSS data (Supplementary Figure S2E). Since the Li *et al.* TSS atlas was also based on CAGE, there was high agreement of precise TSS locations with our set (Supplementary Figure S2E, bottom), while Eser *et al.* TSSs showed a systematic downstream shift in TSS positions (Supplementary Figure S2E, top). This is not surprising given the random shearing and RNA length selection in the RNA-Seq protocol, which makes RNA-Seq prone to underestimating mRNA lengths. See Supplementary Table S2 for details on all TSSs and their relation to PomBase, Li *et al.* and Eser *et al*.

Together, these observations (nascent RNAs and steady-state RNA approaches for locating TSSs, sequence content, nucleosome occupancy and histone modifications) strongly indicate that CAGE-defined TSSs are more accurate than current PomBase annotation, and greatly expands previous systematic attempts at locating TSSs. Thus, our data do not only increase the number of known TSSs, and thereby promoters of *S. pombe*, but additionally refines the positional accuracy of existing genome-wide data sets. The latter is important for any study relying on nucleotide-resolution TSS-anchored analysis and visualization; including profiling of histone marks, transcriptional activity, initiation and elongation. Our data thereby constitute a valuable new source of information for investigating *S. pombe* transcriptional regulation and chromatin biology.

S. pombe promoters lack broad TSS distributions but a subset support bidirectional transcription initiation

One of the main advantages of *S. pombe* compared to *S. cerevisiae* as a model organism is its higher similarity to mammalian genomes in terms of gene regulation and chromatin biology. It was therefore important to assess whether core promoter features in mammalian genomes also were present in *S. pombe*. In particular, we compared the shape of TSS distributions, the prevalence of bidirectional transcription initiation and the relation between TSSs and nucleosomes in *S. pombe* versus mammalian genomes.

In mammals, promoters can be distinguished by their distribution of TSS into 'broad' and 'sharp' classes. Sharply defined TSS distributions, where CAGE tags are focused on one or a few nearby nucleotides, are associated with TATA boxes and cell/tissue-specific transcription, while broader TSS distributions are often overlapping CpG islands and are associated with more ubiquitous expression (reviewed in (60)). Notably, in mammals, broad TSS distributions are more common than sharp. We measured the broadness of nearby TSSs as the width of CAGE TCs (see Supplementary Figure S1A) in *S. pombe* and a set of representative mammalian CAGE libraries produced and processed in the same way as the *S. pombe* libraries presented here (five pooled *M. musculus* lung CAGE libraries (45)).

While both *S. pombe* and *M. musculus* had CAGE TC widths spanning from a single to hundreds of bps, *S. pombe* did not show the same bimodal distribution of TC widths found in mouse; in particular, 30–50 bp wide TCs were common in *M. musculus* but not in *S. pombe*. This observation was correlated with the underlying DNA sequence content: while *M. musculus* TC width, as previously reported, was positively correlated with CpG dinucleotide content (Figure 3A), this relation was not present in *S. pombe*, perhaps related to the allover low CpG dinucleotide content in *S. pombe* promoters (Supplementary Figure S3A). Similarly, in *M. musculus* we observed a clear tendency for stronger TATA-box matches upstream of the peaks of sharp TSS distributions (width 1–5 bp); this pattern was absent in *S. pombe* (Figure 3B); on the other hand, *S. pombe* TCs had, regardless of their width, average TATA match scores as high as sharp *M. musculus* TCs. This is likely partially due to the overall higher T/A content in *S. pombe* promoters. INR motif scores showed no relation to TC width in either species (Supplementary Figure S3B). Overall, our results imply that *S. pombe* does not display broad TSS distributions as frequently as mammals, which may be due to the overall higher TA content and, in extension, higher TATA-box content.

Recent studies have shown that transcription in mammals initiates at both edges of the NDR, producing mRNAs in the sense direction and typically short and exosome-sensitive RNAs in the reverse direction. In mammals, such

reverse strand transcripts have been termed PROMPTS or uaTSS (60–62). Similar observations have been made in *S. cerevisiae*, where the transcripts are called CUTs or SUTs (reviewed in (61)), where CUTs, but not SUTs, are primary exosome targets. In *S. pombe*, the presence of PROMPTs has been debated: a study based on tiling-arrays detected PROMPT-like transcripts (62), while a recent study using nascent RNA sequencing (41) indicated that PROMPTs are uncommon. A third study showed that transcription of PROMPTs may occur but is repressed by the Spt5 protein (63).

To investigate this question in light of our CAGE TSS data, we re-analyzed recent nascent RNA sequencing data (NET-Seq, PRO-Cap and PRO-Seq) (41,42). When anchoring upon all CAGE-defined TSSs, there was no overall pattern of bidirectional transcription in any of the three assays (data not shown), supporting the conclusion drawn in (41). However, through manual observation, we found many cases where a CAGE-defined TSSs had a corresponding nearby bidirectional TSS (for example, see Supplementary Figure S2E). To see if such cases were supported by nascent transcription assays, we identified all CAGE-defined TSSs that also had upstream CAGE signal on the opposite strand within 250 bp. Because the *S. pombe* genome is gene-dense, we filtered cases having one or more additional annotated TSS within 500 bp on any strand.

Many of these CAGE TSSs showed reverse strand signal in PRO-cap, NET-seq and PRO-seq consistent with transcription initiation at the edges of the NDR, mirroring the properties of PROMPTs in mammals, but the signal was inversely correlated with the amount of sense transcription in the NDR, as measured by NET-Seq (Figure 3C and Supplementary Figure S3C). A possible explanation for this observation is that read-through transcription from upstream genes on the same strand suppresses transcription initiation on the reverse strand, although the mechanism is unclear. Indeed, we generally observed stronger PROMPT transcription at lowly expressed forward strand CAGE-defined TSSs near or upstream of annotated promoters, while highly expressed forward strand TSSs in coding regions or 3′-UTR only showed weak PROMPT transcription (Supplementary Figure S3D).

In conclusion, the above analyses indicate that *S. pombe* TSSs share some, but not all mammalian properties: *S. pombe* promoters do not exhibit a clear distinction between broad and sharp TSS distributions, *S. pombe* TSSs are placed next to nucleosome boundaries but only a subset of them seem to support bidirectional initiation. Such cases typically corresponded to TSSs with limited read-through from other transcriptional units.

Identification of stress-responding TSSs

In the above analysis, we treated all TSSs equally, even though they may be ubiquitously expressed across conditions, specifically expressed in one condition, or somewhere between these two extremes. MDS analysis showed that libraries subjected to the same type of stress and medium clustered tightly together, indicating low biological variance within groups and consistent differences between groups (Figure 4A). Nitro-EMM libraries, followed by Heat-YES, appeared the most different from Ctrl-EMM. Interestingly, we observed a clear difference between control libraries growing on the different growth media (Ctrl-EMM and Ctrl-YES), which was of comparable magnitude to the difference between Ctrl-YES and H_2O_2-YES. Thus, YES medium appears to alter the TSS landscape to an extent comparable to some stressors.

To identify individual TSSs that showed statistically significant changes in expression, we analyzed differential expression on TSS-level using the linear model framework implemented in the R-package limma (47) (Supplementary Figure S4A–F), since this allows for the comparison of different conditions while taking into account the different media (e.g. comparing Heat-YES to Ctrl-YES, Nitro-EMM to Ctrl-EMM, etc.). Because of the high sequencing depth combined with the low variance within groups, we were able to detect even weak levels of differential expression across all ranges of TSS expression. However, even though subtle changes could reliably be detected, we reasoned that changes with a substantial fold change were the most biologically relevant. Thus, we only considered TSSs with an absolute $\log_2$ fold change ≥ 0.5 (corresponding to ∼40% higher or lower expression between conditions) and $FDR < 0.05$ to be significantly differentially expressed between two or more states. We defined four sets of differentially expressed TSSs (Figure 4B, left schematic shows pairwise comparisons stemming from the linear model) responding by either increasing or decreasing expression: $\text{Nitro}_{up}/\text{Nitro}_{down}$ for nitrogen starvation, $\text{Heat}_{up}/\text{Heat}_{down}$ for heat shock, H_2O_{2up}/ H_2O_{2down} for hydrogen peroxide stress and $\text{YES}_{up}/\text{YES}_{down}$ for YES media (Supplementary Tables S3 and S7; Supplementary Table S4 shows all pairwise comparisons between groups). The results recapitulated the MDS plot: the Nitro_{up} and Nitro_{down} sets had the highest number of differentially expressed TSSs, followed by $\text{Heat}_{up}/\text{Heat}_{down}$, $\text{YES}_{up}/\text{YES}_{down}$ and H_2O_{2up}/H_2O_{2down} (Figure 4B) Notably, while $\text{YES}_{up}/\text{YES}_{down}$ and $\text{Nitro}_{up}/\text{Nitro}_{down}$ had roughly as many up- as downregulated TSSs, $\text{Heat}_{up}/\text{Heat}_{down}$ and H_2O_{2up}/ H_2O_{2down} showed a substantially higher number of upregulated versus downregulated TSSs.

The three stress conditions shared many differentially expressed TSSs, (Figure 4C and D), indicating that a substantial part of stress response was generic. Notably, almost half (49%) of the 6375 detected *S. pombe* TSSs were differentially expressed in at least one stress condition (excluding YES response). The YES-response was generally composed of distinct TSSs compared to the stress response.

We next investigated how the differential expression results related to established fission yeast stress biology using two different approaches. First, we related our results to Gene Ontology (GO) terms, representing systematic gene function annotation based on a large number of studies. Since GO-term annotations are defined at gene level, for this analysis, we measured gene expression by summing CAGE expression contribution from all CAGE-defined TSSs within a gene on the same strand as in (64), and used these expression values to perform gene-wise tests for differential gene expression in the same way as the TSS-level expression analysis above (Supplementary Figure S5A–J). Most differentially expressed gene sets were enriched for

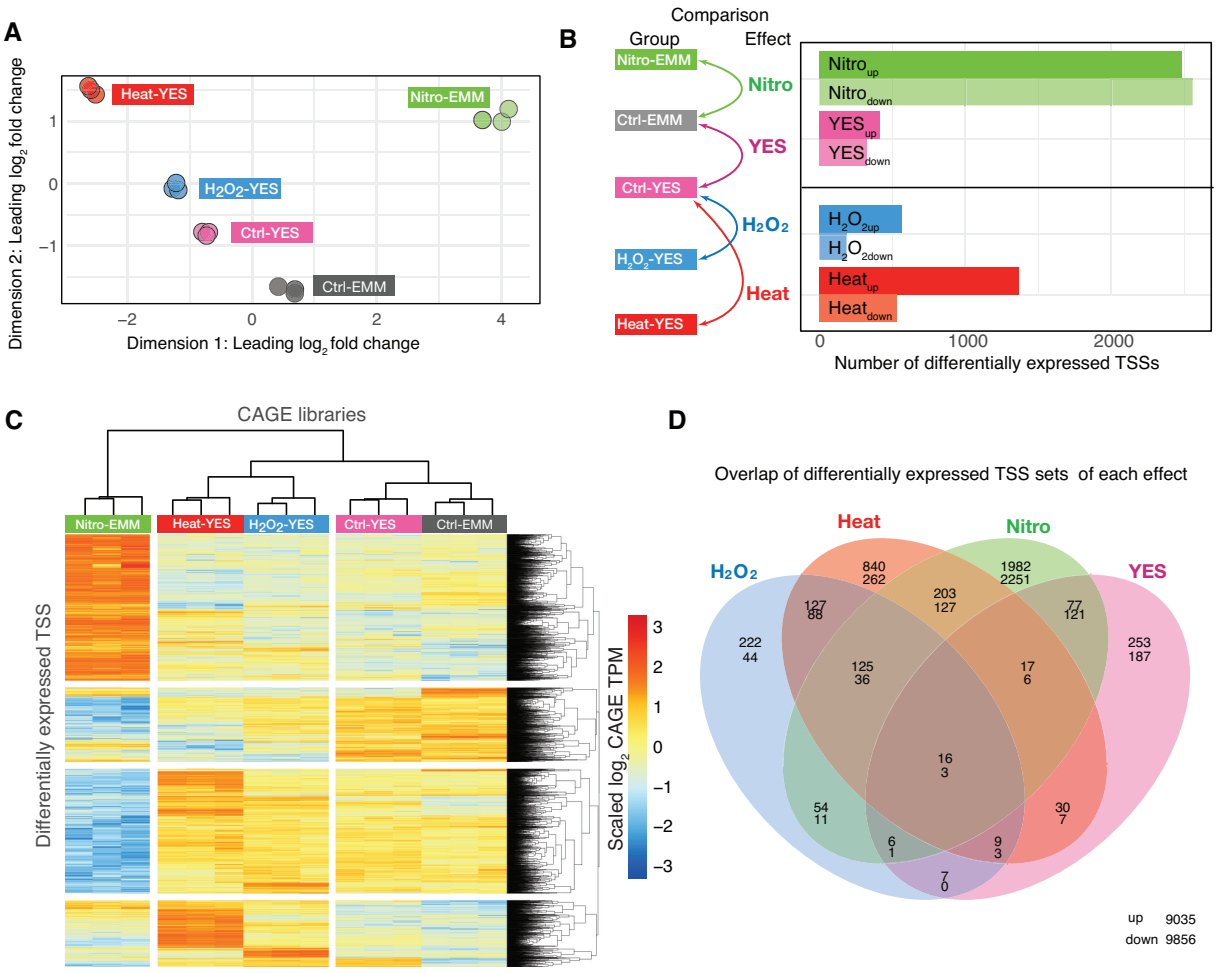

Figure 4. Differential expression of TSS in response to stress and media change. (**A**) Multi-dimensional scaling (MDS) plot of CAGE libraries. MDS *X*-and *Y*-axes show the first two MDS dimensions. Each point corresponds to a CAGE library, colored by condition. Axes are scaled as leading $\log_2$ fold changes; the root-mean-squared average of the $\log_2$ fold changes of the top 1000 genes best separating each sample. (**B**) Differential expression analysis on TSS level. Left: schematic overview of what pairwise comparisons within the linear model that gave rise to differentially expressed TSS sets of interests, using Ctrl-EMM as a baseline. Right: bar plot shows the number of differentially expressed TSSs in each set, split by direction of change. TSSs were considered significant if $FDR < 0.05$ and $|\log_2 \text{fold change}| > 0.5$, using limma-voom. (**C**) Expression patterns of differentially expressed TSSs. Heat map rows correspond to all differentially expressed TSSs in panel B. Columns correspond to all CAGE libraries, as indicated by color. Heat map shows the relative CAGE expression ($\log_2$(TPM) scaled to unit mean/variance across rows) for a given TSS and library. (**D**) Overlap of differentially expressed TSS between treatments. Venn diagram shows the overlap between sets of differentially expressed TSSs as defined in panel B. Number pairs within Venn diagram areas show the number of up- and downregulated TSSs (upper and lower number, respectively). Number in lower right corner are non-differentially expressed TSSs.

the expected GO terms (Supplementary Figure S5K shows the 10 top GO terms for each gene set, Supplementary Table S5 shows all terms), i.e. H_2O_{2up} was enriched for 'cellular response to oxidative stress', $Heat_{up}$ was enriched for 'protein folding', $Nitro_{up}$ was enriched for 'autophagy' and YES_{down} was enriched for 'thiamine biosynthetic process'. We observed an enrichment of GO-terms related to meiosis in genes upregulated after nitrogen starvation (e.g. 'conjugation with cellular fusion' and 'meiosis 1'). This is expected, as sexual development in *S. pombe* is intimately linked to nitrogen starvation (16,65,66), and nitrogen starvation in EMM2 medium is particularly efficient in inducing this response. The combined activation of genes related to sexual development and stress response explains the overall larger number of differentially expressed TSSs/genes in $Nitro_{up}$/$Nitro_{down}$ compared to the other stressors.

Second, we compared the differential expression results in detail with two key papers on fission yeast stress and mating response. Chen *et al.* (4) used microarrays to profile the stress response across a wide range of stimuli to define sets of CESR and SESR genes. The study included peroxide stress and heat shock as stressors similar in duration and strength to the samples used in our study. They reported 290 and 274 genes upregulated > 3-fold by peroxide stress and heat shock, respectively, compared to our 265 H_2O_{2up} and 640 $Heat_{up}$ genes. This difference in overall numbers between the studies likely reflects both the methods used and statistical tests and cutoffs for calling differential expression. CAGE H_2O_2 fold changes for peroxide-induced genes identified in (4) were consistently higher than for any other treatment; the same was true for heat-induced genes identified in the same study, and in both cases all

genes except two showed a congruent direction of change (Supplementary Figure S6A). Similarly, there was a substantial overlap between CESR-upregulated genes defined in (4), and upregulated genes in our sets: 38% (26/68) CESR genes were in H_2O_{2up}, $Heat_{up}$ and $Nitro_{up}$ gene sets, while only seven CESR-upregulated genes were not present in any of our differentially upregulated sets (Supplementary Figure S6B). The overlap with CESR-downregulated genes was smaller, although only 4 genes showed a different direction of change (Supplementary Figure S6C).

The same analysis was made to compare our results to previous microarray profiling of expression during meiosis and nitrogen starvation (15). Mata *et al.* profiled transcripts up to 12 h after pheromone stimulation of nitrogen-starved cells, compared to our 16 h time point of nitrogen starvation without pheromone stimulation. On a genome-wide scale, Mata *et al.* observed around 2000 genes responding at any point of the time course, while we observed 2004 nitrogen starvation-responsive genes (Supplementary Figure S5). Mata *et al.* defined sets of genes showing continuous or delayed activation during the time course. Consistent with the known co-regulation of genes in response to nitrogen starvation and pheromone stimulation (16), CAGE expression showed a corresponding upregulation of genes across both those sets (Supplementary Figure S6D). Only three genes were downregulated in the CAGE data, presumably reflecting the different experimental setup in the two studies.

Thus, as a summary, both GO and comparisons to specific studies showed that our data is consistent with previously established expression patterns in fission yeast stress response.

Since we found that almost half of all TSSs were changing due to one or more stressors, we also wished to identify TSSs and genes that were not changing in expression across any stress conditions. To select such TSSs or genes we required an $FDR > 0.05$ in any comparison between conditions (using an F-test against a $\log_2$ fold change different from 0), resulting in 19% (2530) of TSSs and 16% (950) of genes. These genes were enriched for GO-terms related to protein transport and the Golgi-apparatus, e.g. 'intracellular protein transport' and 'ER to Golgi vesicle-mediated transport' (Supplementary Figure S7A). TSSs/genes showing no differential expression between a wide range of conditions would be useful as candidates for 'housekeeping TSSs/genes', and thereby good references, e.g. used for qPCR controls. We reasoned that ideal candidates should show low variance of expression across and within conditions. Because variance is correlated with expression strength, we required candidates to show low variance of expression given their expression levels, which resulted in a list of 93 TSS and 49 housekeeping gene candidates (Supplementary Tables S6 and S7). We also looked specifically at a set of commonly used housekeeping genes: *act1*, *adh1*, *atb2*, *cdc2*, *gad8*, *leu1*, *ura4*. While all these genes showed low variance of expression, some of them also showed differential expression between conditions. In particular, all housekeeping genes except *act1* and *cdc2* exhibited a high fold change in response to nitrogen starvation. Only *cdc2* did not show differential expression in any comparison (Supplementary Figure S7B and C). This finding implies that careful selection of reference genes/TSSs is important.

Stress responding TSSs provide insights into *S. pombe* gene regulation

An important question is how the various stress responses identified above are regulated. Because CAGE data enables both accurate localization of TSSs, and precise quantification of TSS usage across conditions, it is ideally suited for prediction of transcription factor binding sites (TFBSs) responsible for stress-specific gene regulation in upstream promoter regions.

We used HOMER (49) to identify enriched DNA motifs in the promoters around TSSs up- or downregulated by each stressor (defined as the −300 to +100 bp around TSS peaks). Because there is no comprehensive resource of transcription factor motifs in fission yeast, we performed the analysis *de novo* (identifying over-represented motifs with no prior information). To determine whether any of the *de novo* motifs resembled known motifs, we compared them to HOMER's database of motifs from *S. cerevisiae*.

As with GO terms, we observed motifs enriched in specific types of stress as well as motifs common across all stressors. In many cases, the enriched patterns reflected known *S. pombe* or *S. cerevisiae* biology (see Figure 5 for a summary and full analysis in Dataset S1). $Heat_{up}$, H_2O_{2up} and $Nitro_{up}$ promoters all shared enrichment for the CST6 motif. CST6 is a homolog of Atf1 which is widely regarded as the central regulator of CESR (4,67). $Heat_{up}$ promoters were highly enriched for a HSF1-like motif, while H_2O_{2up} promoters were enriched for a PAP1-like motif (*S. cerevisiae* homolog of *YAP1*). Heat shock factors and PAP1 are well-known transcription factors associated with the specific response to heat shock and hydrogen peroxide stress, respectively, in both *S. pombe* and *S. cerevisae* (68,69).

$Nitro_{down}$ promoters were enriched for the SFP1 motif, while H_2O_{2down} was enriched for the similar SUM1 motif (while SFP1 was not found *de novo* in $Heat_{down}$, a complimentary enrichment analysis using known *S. cerevisiae* motifs showed that SFP1 was enriched in $Heat_{down}$, H_2O_{2down} and $Nitro_{down}$). SFP1 was previously implicated in regulating ribosomal gene expression (70), supporting observed enrichment of the 'ribosome biogenesis' GO-term in H_2O_{2down}, $Heat_{down}$, $Nitro_{down}$ genes (Supplementary Figure S5H). As discussed above, $Nitro_{up}$ were enriched for genes associated with sexual development in *S. pombe*. In agreement, the Ste11 motif (*AZF1* homolog) (16) was highly enriched in $Nitro_{up}$ as well as a GATA binding site (GAT1 homolog). Ste11 and the GATA transcription factor Gaf1 have previously been shown to be regulated in a coordinated fashion in *S. pombe* (71).

YES_{up} and YES_{down} were enriched for different motifs when compared to the three types of stresses, reflecting an overall different regulation of this response. We are not aware of any comprehensive study of TFBSs involved in the response to YES-media, but our analysis suggests a possible role for several novel *S. pombe* TFBSs similar to known TFBSs in *S. cerevisae*. For example, YES_{up} were enriched for a PDR1 motif and a general forkhead motif. PDR1 is annotated as specific to *S. cerevisiae*, while at least four forkhead-

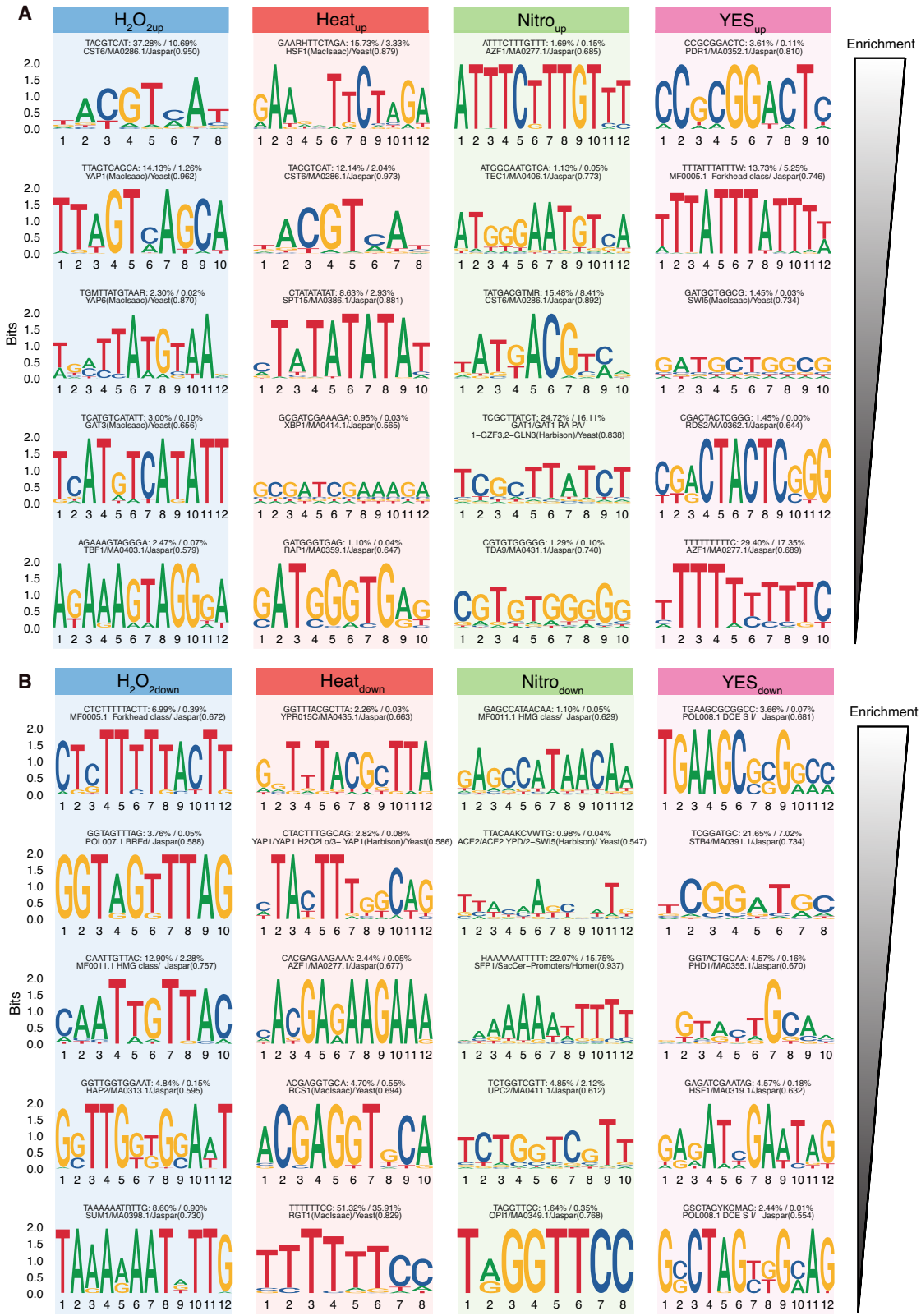

Figure 5. *De novo* DNA motif enrichment in promoters of differentially expressed TSSs. (**A**) Enriched motifs in promoters of upregulated TSSs. Columns show the top five enriched *de novo* motifs in the promoter regions of each set of differentially expressed TSSs (as in Figure 4B), where the top motif is the most enriched. Motifs are shown as sequence logos, where *Y*-axis shows bit score and *X*-axis shows nucleotide positions in bp. Text above logos shows consensus sequence, foreground/background frequency and closest matching known *Schizosaccharomyces cerevisae* motif (score in parenthesis indicated quality of match). (**B**) Enriched motifs in promoters of downregulated TSSs. Panels are arranged as in panel A, but show enriched motifs for promoters of downregulated TSSs.

domain-containing genes exist in *S. pombe* (*fhl1, fkh2, mei4 and sep1*), which may have a role in cell adjustment to rich media. YES_{down} promoters were enriched for GCR1, STB4 and PHD1 motifs. The GCR1/2 genes have not been shown to have *S. pombe* homologs, but regulate the glycolytic pathway in *S. cerevisiae* (72).

Single nucleotide variation around *S. pombe* TSSs

TFBSs underlying the enriched motifs above were primarily detected in the −50 to −150 bp region relative to TSSs, suggesting that much of the regulatory signals for stress response may be located in this region. To investigate how this over-representation related to genetic variation across *S. pombe* populations, we investigated the overlap of CAGE-defined TSSs with available SNVs generated from comparison of multiple global *S. pombe* strains (50) (Figure 6A). Overall, and consistent with (50), we observed a higher number of SNVs in intergenic regions (i.e. upstream of TSSs) compared to intragenic regions (downstream of TSSs). However, the precise location of the CAGE-defined TSSs revealed additional SNV patterns.

First, downstream of TSSs, SNV abundance was highly correlated to nucleosome occupancy (as defined by MNase-Seq data, see Figure 2F) and anti-correlated to CG dinucleotide content (Figure 6A). It is interesting to note that while CG dinucleotides in general are hyper-mutable, the TA-rich regions within the first 6–7 nucleosomes downstream of the TSS showed higher average SNVs than in CG-rich linker regions. A possible underlying mechanism is that DNA with high nucleosome occupancy have higher mutation rates because nucleosomes block the access of the DNA repair machinery, as has been observed in human (73).

Second, SNV abundance was highest −400 to −100 bp upstream of TSSs, but rapidly declined toward two minima corresponding to the TATA box position at around −30 and the TSS at +1 (see zoom-in panel of Figure 6A). The low frequency of observed SNVs around the TATA-box and TSS was similar to the SNVs frequency in coding exons. These results indicated that while the regions immediately surrounding the TSS and TATA-box were similar between *S. pombe* strains, the region upstream of the TSS was more variable: this was somewhat unexpected as this region overlapped most predicted TFBSs. To investigate whether this pattern was related to differential expression, we examined whether the regions around differentially expressed TSSs had more or less SNVs versus non-differentially expressed TSSs. Specifically, we modeled the number of SNVs in the −300 to +100 region around the TSSs as a function of (i) their differential expression ($Heat_{up}/Heat_{down}$, YES_{up}/YES_{down}, etc. as in Figure 4B) and (ii) their overlap with gene annotation (annotated TSS, 5′ UTR, coding sequence, etc., as in Figure 2A) using negative binomial regression (Figure 6B). The advantage with this approach is that the effect of TSS location and differential expression status can be separated. Compared to a baseline consisting of non-differentially expressed TSSs near annotated promoters, intergenic TSSs (annotated as proximal or intergenic) had more SNVs, while intragenic TSSs (intron, antisense, 3′-UTR and CDS) had less (in agreement with Figure 6A). Correcting for these gene location effects through the regression model, we found that some differentially expressed sets of TSSs were associated with substantial increases in the number of SNVs. In particular, YES_{up} and YES_{dowm} promoters had an increase in SNVs larger than the effect of being in an intergenic region. H_2O_{2up}, $Heat_{down}$ and $Nitro_{up}$ promoters were also enriched for SNVs, while H_2O_{2down}, $Heat_{up}$ and $Nitro_{down}$ were not enriched or depleted. We speculate that this increase in SNVs in differentially expressed promoters in particular could be related to positive selection or genetic drift of SNVs in these regions (see 'Discussion' section).

Stress-specific activation of alternative TSSs

Alternative TSSs are common in multicellular organisms (29,30,74). Principally, genes may utilize alternative TSSs for at least three reasons: (i) Having multiple TSSs or promoters may increase the overall expression of the gene in question; (ii) If alternative TSSs are located within genes, their usage may produce RNAs that exclude coding exons resulting in truncated proteins, which may have a functional impact. (iii) Multiple TSSs offer higher regulatory flexibility, where two TSS may respond to different stimuli or context due to different (TFBSs) around them.

While several studies have shown context-specific usage of alternative TSSs that may be disruptive and produce truncated proteins in *S. cerevisiae* (e.g. (74–78)), to our knowledge, no systematic attempts to profile alternative TSS usage in the *S. pombe* genome under different environmental conditions have been done. We reasoned that since CAGE can detect and quantify both known and novel TSS, our data set is well suited for detecting differential TSS usage within genes across stress conditions.

First, we investigated the occurrence of multiple TSS inside genes across all conditions. Around 54% of genes had >1 TSSs (Figure 7A), indicating that multiple TSSs per gene is a widespread phenomenon in *S. pombe*. For many genes, a single dominant TSS might be expressed at much higher levels than other TSS(s) within the same gene. We therefore identified the dominant TSS within each gene (defined as the TSS with the highest overall expression across all libraries) and analyzed where dominant and non-dominant TSSs were located respective to annotation (Figure 7B). For genes having >1 TSSs, the dominant TSS most often overlapped the annotated TSS region or 5′ UTRs, much like TSSs within single TSS genes. Non-dominant TSSs were most commonly found in coding regions (CDS) or proximal upstream regions, a pattern consistent with the overall expression of TSSs in the same regions (Figure 2A).

Next, we wanted to investigate how often multi-TSS genes used different TSS under different conditions. For each gene, we analyzed whether any of its TSSs showed a different response across conditions compared to the other TSSs in the same gene (using the diffSplice approach from limma, see 'Materials and Methods' section). Nitro resulted in the highest number of genes showing differential TSS usage, followed by Heat, H_2O_2 and YES, mirroring the overall gene expression results (Figure 7C and D). Overall, 77% percent of multi-TSS genes showed differential TSS expression compared to the gene's other TSSs in at least one condition. It is thus common for genes to have at least two

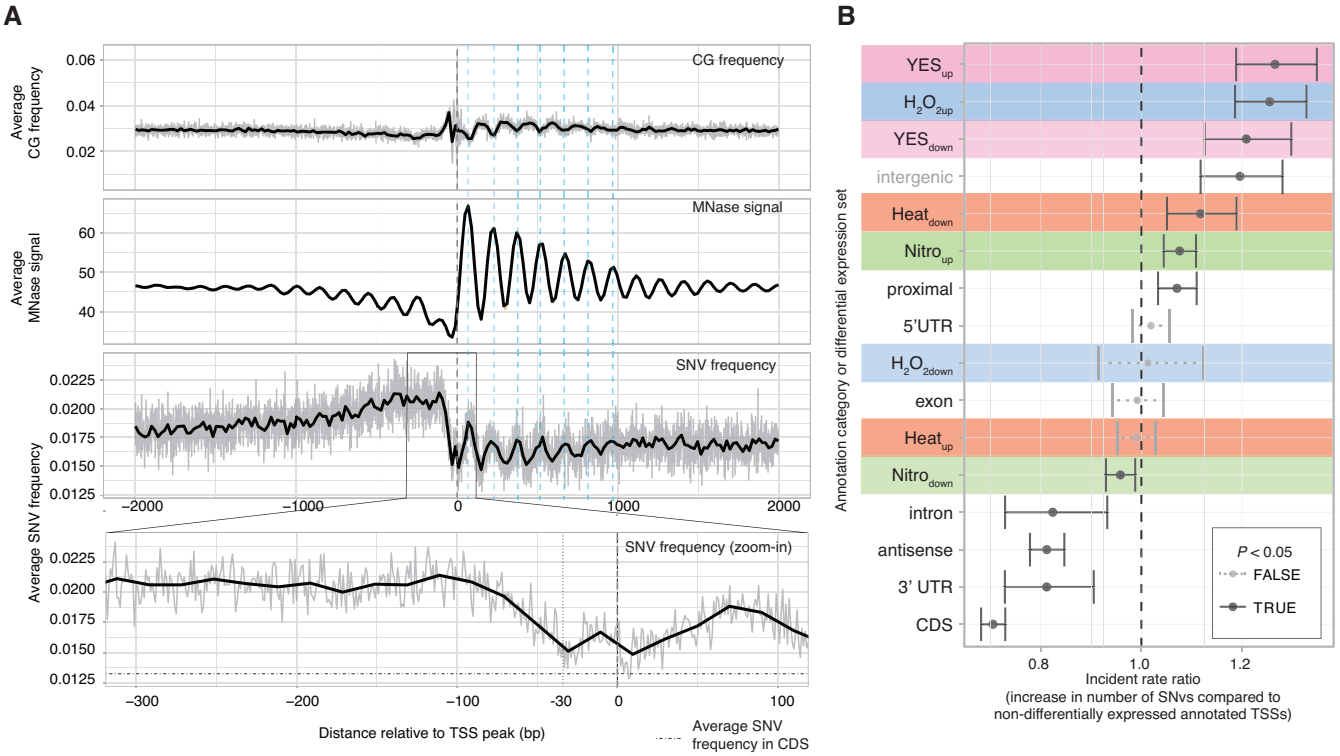

Figure 6. SNV frequency around CAGE-defined TSSs. (**A**) SNVs around TSSs compared to CG-content and MNase-Seq signal. *X*-axis in all panels shows distance from the CAGE-defined TSS peak in bp. Top panel shows average CG di-nucleotide frequency (gray lines show bp-level averages and black lines show a trend line, as in Figure 2E). Middle panel shows average MNase-Seq signal, where larger values correspond to positioned nucleosomes (similar to Figure 2F). The lower panel shows average number of SNVs, with zoom-in panel showing the −300 to +100 promoter region (gray lines show bp-level averages and black lines show a loess regression trend line). Vertical-dashed teal lines indicate regions with high nucleosome occupancy. For the zoom-in panel, horizontal dot-dashed line indicate the average number of SNVs in coding exon regions as a reference, vertical-dashed lines indicates the −30 position (typical TATA box position) and the 0 position (the TSS peak). (**B**) Model of number of SNVs in promoters across annotation categories and differential expression sets. *Y*-axis shows the different modeled effects, composed of different annotation categories (as in Figure 2A) and the differential expression sets (as in Figure 4B, also indicated by background colors). *X*-axis shows the incidence rate ratio: the estimated fold-change increase in the number of SNVs, compared to a non-differentially expressed TSS near a PomBase-annotated TSS. Points indicate estimates for the incident rate ratio for every modeled effect, whiskers indicate 95% confidence intervals. Whisker color indicates whether the *P*-value associated with an incident rate ratio is <0.05.

TSSs that are regulated differently in *S. pombe*. There was no clear preference for differential TSS usage in any specific part of genes (Figure 7D) when comparing to the overall number TSSs in those categories (Figure 7B), i.e. there was no enrichment (or depletion) of differential TSS usage in any particular part of genes compared to other TSSs. The largest number of differentially used TSSs was therefore observed in annotated promoter regions or 5′-UTRs. This pattern was consistent across all conditions.

Many genes showed both differential expression and differential TSS usage as defined above (Supplementary Figure S8A), indicating that these processes are coupled. We observed the same overall pattern in the top most expressed TSSs within genes (only TSSs making up at least 10% of total gene expression in at least three libraries) (Supplementary Figure S8B–D).

In several cases, differential TSSs usage detected by CAGE verified previous single-gene studies. For example, we detected two TSSs in the 5′ UTR of the *sod1* (Superoxide Dismutase 1) gene where most of the upstream TSS showed little change between conditions, but the downstream TSS (∼1 kb downstream of the annotated sod1 TSS) showed high expression in all EMM2 media conditions and low expression in YES-media conditions (Figure 7E). Two *sod1* mRNA isoforms were previously identified (79): a shorter transcript that was gradually replaced by a larger transcript when cultures growing in YES entered stationary phase. The lengths of both transcripts detected in their Northern blot corresponded to the distance between the two TSSs detected by CAGE, suggesting that a switch from EMM2 to YES can create a similar change (Figure 7E, inset). The *gpa1* locus is another example of differential expression of alternative TSSs (Supplementary Figure S8E)

We were also interested in the specific cases where differential TSS usage could lead to an altered protein product. We reasoned the most dramatic effect on the final protein product of a gene would be the exclusion or truncation of protein domains by the use of alternative TSSs. We identified TSSs that were located within or downstream protein domains defined in PomBase. Usage of these 'disruptive' TSSs would give rise to either domain truncation or exclusion. We found that 64% of CDS TSSs, 56% of intronic TSSs and 75% of 3′-UTR TSSs were classified as disruptive. The average expression level of disruptive TSSs was low, and dis-

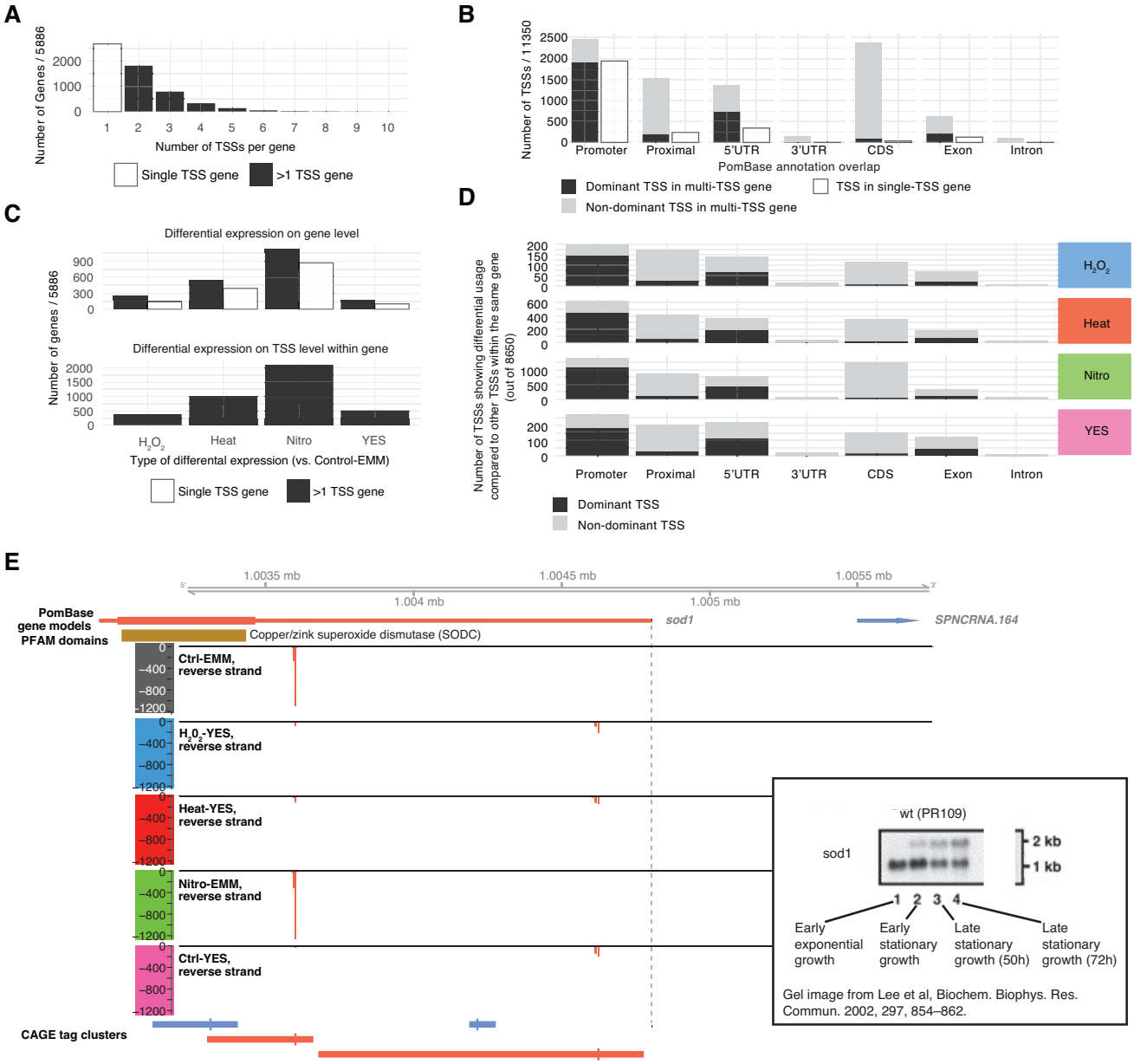

Figure 7. Differential TSS usage within genes. (**A**) Number of TSSs per gene. *X*-axis indicates the number of TSSs per gene and *Y*-axis the number of genes. Colors indicate single-TSS genes (white) and multi-TSS genes (black). (**B**) TSS structure compared to annotation and expression. *X*-axis shows the annotation categories of intragenic TSSs (as in Figure 2A), bar plots show the number of TSSs in each category, sub-divided into dominant (most highly expressed, black) and non-dominant TSSs within genes having >1 TSS (gray) and TSSs within single-TSS genes (white). (**C**) Number of genes showing differential TSS usage compared to differential gene expression. The top panel shows the total number of differentially expressed genes for each effect; genes are divided into those having one or more TSSs. The bottom panel shows the number of genes having at least one TSS with differential expression compared to the rest of TSSs within the gene. *X*-axes indicate effects and *Y*-axis indicates the number of genes. (**D**) Differential TSS usage across different annotation categories. Bar plots show the number of TSSs having differential expression compared to the rest of TSSs within the same gene, split by effect (rows, indicated by color), and PomBase annotation overlap (columns). (**E**) Genome browser example of differential TSS usage in the *sod1* locus. Plot is organized as in Figure 1C. Inset shows Northern blot experiment on *sod1* RNA from ref (79) with added annotations. See main text for details.

ruptive TSSs generally made low contributions to total gene expression (Figure 8A). We detected 52 dominant disruptive TSSs, and 478 disruptive TSSs which contributed more than 10% of the expression of their host gene in at least three libraries. Of these TSSs, 356 were differentially used in at least one condition (Figure 8B). This indicates that while alternative TSS usage affecting protein domain inclusion is relatively rare, there are clear cases of this phenomenon.

The *cds1* gene represents an example of a domain-disruptive TSS (Figure 8C). *cds1* had two highly expressed TSSs: one in the 5′ UTR which was active across all conditions, and one in the fourth protein coding exon, expressed only in Nitro-EMM and Heat-YES. The latter TSS would produce an RNA that cannot encode the forkhead domain encoded in the upstream exons. A previous study showed that this domain is important for the activation of

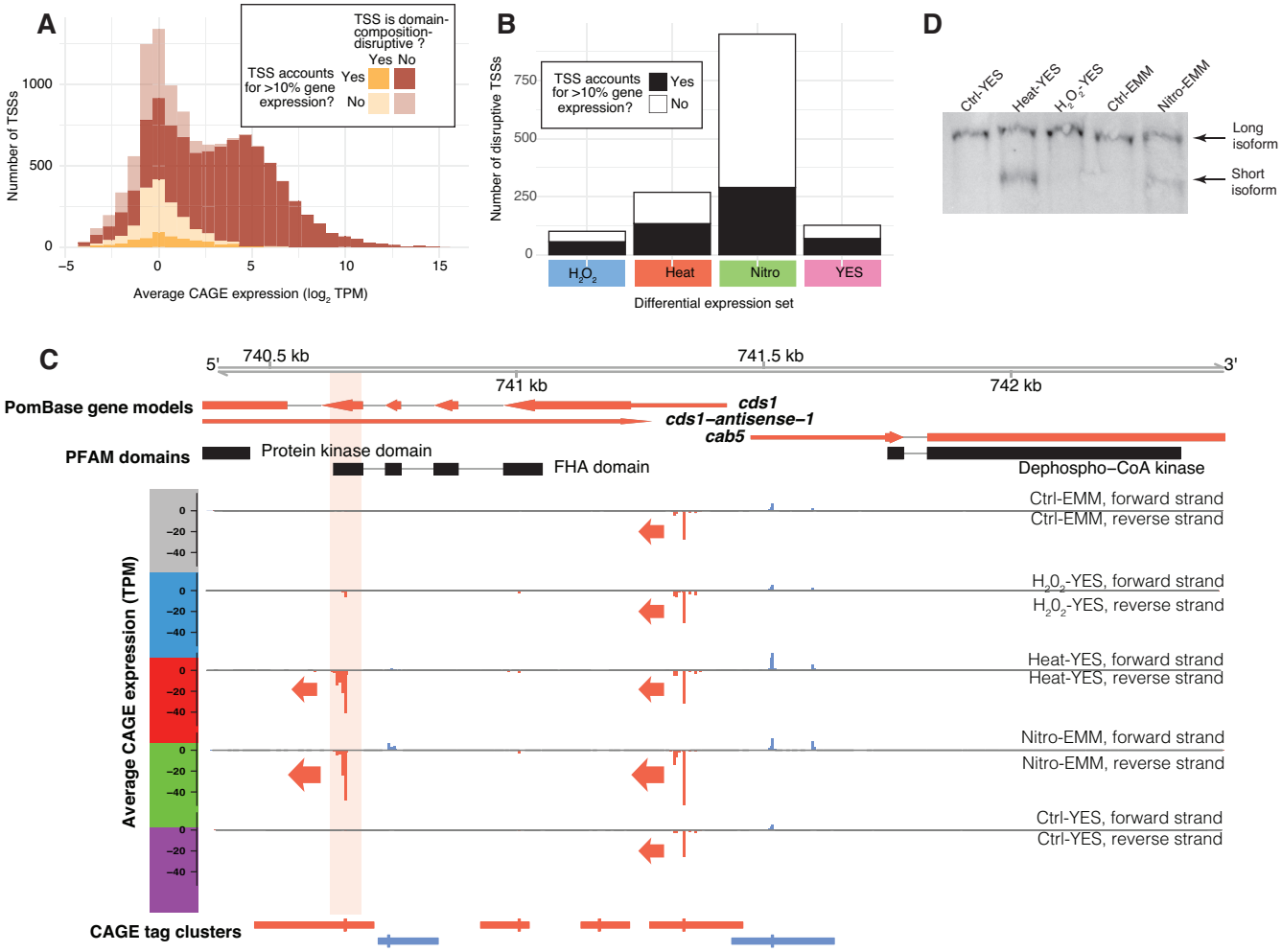

Figure 8. Candidates for alternative TSSs affecting domain composition. (**A**) Expression levels of domain composition-disruptive TSSs (TSSs in or downstream of protein domains): stacked histogram of TSS expression; *X*-axis show $\log_2$(TPM), *Y*-axis show number of TSSs. TSSs are split depending on whether the TSS disrupts domain structure and whether it contributes more than 10% of total gene expression in at least three libraries, as indicated by legend. (**B**) Number of differentially used alternative TSSs potentially affecting domain composition: *X*-axis shows the different conditions. *Y*-axis shows the number of differentially used TSSs with potential to disrupt domain composition. Color indicates whether the TSS contributes more than 10% of total gene expression in at least three libraries (See also Supplementary Figure S8B–D). (**C**) Genome browser example of differential TSS usage in the *cds1* locus. Plot is organized as in Figure 1C, but also shows PFAM domains. TSSs downstream of, or overlapping, PFAM domains are highlighted with light red background. See main text for details. (**D**) Northern analysis of *cds1* using a probe in exon 5. The same RNA samples used for constructing the CAGE libraries were analyzed by northern blotting. The blot was hybridized to a probe in exon 5 of *cds1*. In addition to the full-length transcript present in all samples, the shorter *cds1* isoform, originating from the downstream TSS identified in panel C, is only evident in the Heat-YES and Nitro-EMM samples (lower bands).

Cds1 through binding to Mrc1 (80). We validated the Nitro-EMM and Heat-YES-specific expression of the shorter isoform using northern blotting (Figure 8D).

An important caveat is that it is not possible based on CAGE data alone to determine if the transcript produced by an alternative TSS is translated into a stable protein product. While CAGE can be used as a hypothesis generator, protein-based experiments are needed to verify candidates for protein-altering alternative TSSs. Illustrating this caveat, a western blot for C-terminally tagged Cds1 after heat shock and nitrogen starvation did not detect a protein product corresponding to the Heat-YES/Nitro-EMM-specific downstream TSS discussed above, despite the high RNA levels detected by CAGE (Supplementary Figure S9A). *fus1* and *arg11* genes are additional examples of domain-disruptive TSS (Supplementary Figure S9B and C).

DISCUSSION

Here we have used CAGE to define a comprehensive atlas of TSSs in *S. pombe* and their activity across a range of different growth conditions and stresses. We then compared our TSSs atlas to existing *S. pombe* annotation from PomBase, which is primarily based on RNA-Seq. We show that our TSS atlas can refine the location of known TSSs, including TSSs for protein coding genes, as well as more lowly expressed non-coding transcripts. This is in line with previous work (48), showing pervasive transcription of many unstable transcripts, including many anti-sense transcripts as exemplified in Supplementary Figure S1E.

Accurate TSSs are essential for studying chromatin biology, for which *S. pombe* is a widely used model organism. CAGE-defined TSSs had expected core promoter motifs at expected spacing, and showed a much better agreement with MNase-Seq and ChIP-Seq data for relating transcription initiation with nucleosome positioning and histone modifications respectively. We found that the typical peak distribution of TSSs in *S. pombe* promoters was considerably sharper than that of mammals, with a higher incidence of TATA boxes, perhaps due to the higher TA content in the *S. pombe* genome. As in vertebrates, TSSs are located next to nucleosome boundaries; on the other hand, only a subset of promoters seemed to support initiation of bidirectional transcription. This may be an effect of the closely located genes in *S. pombe*, as the strongest bidirectional TSSs were observed in cases with little read-through from other genes.

The only genome-wide TSS experiment previously reported for *S. pombe* is, to our knowledge, the single replicate CAGE experiment presented in (44). This study showed that TATA and INRs motifs in *S. pombe* are more similar to their vertebrate counterparts than *S. cerevisiae*, which our data also confirms. This study indicated that CAGE-based TSSs had much lower frequencies of TATA/INR patterns if they were not proximal to PomBase TSSs; our data shows that even novel TSSs have TATA/INR patterns. This is likely an effect of our higher number of mapped CAGE tags and lower rRNA contamination versus Li *et al.* (Supplementary Table S1). Lastly, Li *et al.* found a novel motif upstream of a small subset of very sharply defined TSS distributions. Since our motif analysis was based on comparisons between environmental states rather than genomic background, it is not surprising we did not find the same pattern.

Our differential expression analysis on gene- and TSS-level shows that nearly half of TSSs are affected by stress response, with expression consistent with previous microarray studies (4). The overlap is noteworthy, since our study only measures RNA levels at a single time point following stress application, and thus cannot detect genes showing complex responses over time. Interestingly, we also observed a large transcriptional effect of YES- compared to EMM2-media, comparable in overall magnitude to some stresses. A similar phenomenon was reported in (48). The implication of this is that the two most commonly used growth media for *S. pombe* are not interchangeable, and thus, biological conclusions based on experiments performed on YES growing cells cannot necessarily be extrapolated to EMM2 conditions and *vice versa*.

This is also the first dedicated genome-wide study of differentially expressed alternative TSSs in *S. pombe* across stress conditions. We show that differential TSS usage is widespread in the *S. pombe* genome: 54% of genes utilized more than one TSS, and 77% percent of these showed differential alternative TSS usage in at least one condition. The large number of alternative TSSs reported here disagrees with a previous effort to detect genes with alternative TSS in 5′-UTRs using RNA-Seq (48), which identified only 30 genes. Using our data, we could confirm 23 of these cases for 28 genes found in both studies. The discrepancy in numbers is likely due to technical issues: since RNA-Seq is based on random fragmentation, it is challenging to identify distinct 5′-ends of full-length RNAs, as opposed to a dedicated technique like CAGE that specifically selects for them through cap-trapping. It is thus likely that *S. pombe* uses a large repertoire of alternative TSSs. Because multi-exonic genes, required for the generation of alternative splice isoforms, are uncommon in *S. pombe* (only 19.5% percent of genes have >2 exons), our data suggests that alternative TSSs are a greater source of transcript diversity than alternative splicing in *S. pombe*, unlike in mammals where the two processes seem to be used to a similar extent (29,81). We show that most differential TSS usage, where two or more TSSs for the same gene show different transcriptional patterns, involves TSSs located upstream of the annotated TSS or within the 5′-UTR. Differential regulation of such TSSs likely does not change the protein product of the resulting transcript but may confer regulatory flexibility because the two regions can evolve independently.

Binding of transcription factors to sites proximal to TSSs is a key process in gene regulation, and differential expression can be related to the accurate promoter regions by our TSS atlas. Using such regions, we identified likely TFBSs involved in the stress response, ranging from well-known stress-associated transcription factors to possible candidate transcription factors that may be preferentially used in YES or EMM2 media. We reasoned that the regulatory importance of these promoter regions may be related to their evolution. Indeed, we observed a strong conservation of the immediate TSS and TATA-box regions (comparable to that of coding sequence), while regions upstream of the TATA-box were more genetically variable between *S. pombe* populations. A simple explanation for this observed pattern is a lack of negative selection in the predominantly intergenic upstream regions, leading to a higher rate of genetic drift between *S. pombe* strains. However, we also observe that differentially expressed TSSs in many cases show more genetic variability, in particular YES-responding TSSs. Therefore, an alternative model is that while the relatively constrained CDS and core promoter regions are under negative selection, positive selection may be acting on genetic variants affecting transcription factor binding upstream of the TATA-box, thus giving rise to variation in gene regulation between *S. pombe* strains, rather than variation in the structure of genes themselves. Because YES-media is richer than EMM2, we speculate this higher diversity may be related to gradual adaptation to different nutritional environments for the different *S. pombe* strains, while the more acute stress (e.g. heat shock) response patterns may have a higher conservation of their regulatory elements.

While most of the work in this paper has focused on genome-level interpretation of the TSS atlas, we believe the data will have the largest impact as a resource for detailed investigation of individual genes and processes. To make the TSS atlas as easily available as possible to researchers, it is available as a collection of genome browser tracks via the PomBase genome browser.

DATA AVAILABILITY

CAGE raw data, as well as processed TC and bp resolution CAGE TSSs datasets have been deposited in the GEO

database (accession number: GSE110976). The datasets, in summarized form, are also accessible through the PomBase genome browser (Under 'Transcription Start Sites').

Data sets are available in Supplementary Data: compressed archive of complete results of *de novo* and known motif enrichment analysis with Homer. Each differentially expressed set has its own folder, containing a web page named 'homerResults.html' for *de novo* motif analysis and 'knownResults.html' for known motif analysis, as well as position weight matrices for all motifs. For more information on the format see http://homer.ucsd.edu/homer/ngs/peakMotifs.html.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Carsten Skou Nielsen and Peter Brodersen for help with the northern analysis.
Authors' contributions: M.T., A.T., A.A., Y.C., A.S. analyzed data. J.L., M.B. made CAGE libraries. O.N. made *S. pombe* cultures and northern blotting. C.H. performed western blotting. M.T., K.E., A.N., C.W., A.S. interpreted results. M.T., O.N., A.S. wrote the paper with input from all authors.

FUNDING

Lundbeck Foundation (to A.S.); Novo Nordisk Foundation (to A.S.); Villum Fonden (to O.N.); Swedish Research Council (to K.E.); Cancerfonden (to K.E.). Funding for open access charge: University of Copenhagen.
Conflict of interest statement. None declared.

REFERENCES

1. Russell,P. and Nurse,P. (1986) Schizosaccharomyces pombe and Saccharomyces cerevisiae: a look at yeasts divided. *Cell*, **45**, 781–782.
2. Allshire,R.C. and Ekwall,K. (2015) Epigenetic regulation of chromatin states inSchizosaccharomyces pombe. *Cold Spring Harb. Perspect. Biol.*, **7**, a018770.
3. Wilkinson,M.G. and Millar,J.B. (1998) SAPKs and transcription factors do the nucleocytoplasmic tango. *Genes Dev.*, **12**, 1391–1397.
4. Chen,D., Toone,W.M., Mata,J., Lyne,R., Burns,G., Kivinen,K., Brazma,A., Jones,N. and Bähler,J. (2003) Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell*, **14**, 214–229.
5. Chen,D., Wilkinson,C.R.M., Watt,S., Penkett,C.J., Toone,W.M., Jones,N. and Bähler,J. (2008) Multiple pathways differentially regulate global oxidative stress responses in fission yeast. *Mol. Biol. Cell*, **19**, 308–317.
6. Papadakis,M.A. and Workman,C.T. (2015) Oxidative stress response pathways: fission yeast as archetype. *Crit. Rev. Microbiol.*, **41**, 520–535.
7. Sánchez-Mir,L., Salat-Canela,C., Paulo,E., Carmona,M., Ayté,J., Oliva,B. and Hidalgo,E. (2018) Phospho-mimicking Atf1 mutants bypass the transcription activating function of the MAP kinase Sty1 of fission yeast. *Curr. Genet.*, **64**, 97–102.
8. Wilkinson,M.G., Samuels,M., Takeda,T., Toone,W.M., Shieh,J.C., Toda,T., Millar,J.B. and Jones,N. (1996) The Atf1 transcription factor is a target for the Sty1 stress-activated MAP kinase pathway in fission yeast. *Genes Dev.*, **10**, 2289–2301.
9. Toone,W.M., Kuge,S., Samuels,M., Morgan,B.A., Toda,T. and Jones,N. (1998) Regulation of the fission yeast transcription factor Pap1 by oxidative stress: requirement for the nuclear export factor Crm1 (Exportin) and the stress-activated MAP kinase Sty1/Spc1. *Genes Dev.*, **12**, 1453–1463.
10. Sakurai,H. and Takemori,Y. (2007) Interaction between heat shock transcription factors (HSFs) and divergent binding sequences: binding specificities of yeast HSFs and human HSF1. *J. Biol. Chem.*, **282**, 13334–13341.
11. Jia,X., He,W., Murchie,A.I.H. and Chen,D. (2011) The global transcriptional response of fission yeast to hydrogen sulfide. *PLoS One*, **6**, e28275.
12. Fantes,P. and Nurse,P. (1977) Control of cell size at division in fission yeast by a growth-modulated size control over nuclear division. *Exp. Cell Res.*, **107**, 377–386.
13. Shiozaki,K. and Russell,P. (1995) Cell-cycle control linked to extracellular environment by MAP kinase pathway in fission yeast. *Nature*, **378**, 739–743.
14. Hartmuth,S. and Petersen,J. (2009) Fission yeast Tor1 functions as part of TORC1 to control mitotic entry through the stress MAPK pathway following nutrient stress. *J. Cell Sci.*, **122**, 1737–1746.
15. Mata,J., Lyne,R., Burns,G. and Bähler,J. (2002) The transcriptional program of meiosis and sporulation in fission yeast. *Nat. Genet.*, **32**, 143–147.
16. Xue-Franzén,Y., Kjærulff,S., Holmberg,C., Wright,A. and Nielsen,O. (2006) Genomewide identification of pheromone-targeted transcription in fission yeast. *BMC Genomics*, **7**, 303.
17. Lizio,M., Deviatiiarov,R., Nagai,H., Galan,L., Arner,E., Itoh,M., Lassmann,T., Kasukawa,T., Hasegawa,A., Ros,M.A. *et al.* (2017) Systematic analysis of transcription start sites in avian development. *PLoS Biol.*, **15**, e2002887.
18. Rach,E.A., Yuan,H.Y., Majoros,W.H., Tomancak,P. and Ohler,U. (2009) Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biol.*, **10**, R73.
19. Haberle,V., Li,N., Hadzhiev,Y., Plessy,C., Previti,C., Nepal,C., Gehrig,J., Dong,X., Akalin,A., Suzuki,A.M. *et al.* (2014) Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, **507**, 381–385.
20. Schor,I.E., Degner,J.F., Harnett,D., Cannavò,E., Casale,F.P., Shim,H., Garfield,D.A., Birney,E., Stephens,M., Stegle,O. *et al.* (2017) Promoter shape varies across populations and affects promoter evolution and expression noise. *Nat. Genet.*, **49**, 550–558.
21. Morton,T., Petricka,J., Corcoran,D.L., Li,S., Winter,C.M., Carda,A., Benfey,P.N., Ohler,U. and Megraw,M. (2014) Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. *Plant Cell*, **26**, 2746–2760.
22. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A.M., Taylor,M.S., Engström,P.G., Frith,M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
23. Rach,E.A., Winter,D.R., Benjamin,A.M., Corcoran,D.L., Ni,T., Zhu,J. and Ohler,U. (2011) Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLos Genet.*, **7**, e1001274.
24. Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J.L., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M., Itoh,M. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
25. Chen,Y., Pai,A.A., Herudek,J., Lubas,M., Meola,N., Järvelin,A.I., Andersson,R., Pelechano,V., Steinmetz,L.M., Jensen,T.H. *et al.* (2016) Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat. Genet.*, **48**, 984–994.
26. Takahashi,H., Lassmann,T., Murata,M. and Carninci,P. (2012) 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.*, **7**, 542–561.
27. Adiconis,X., Haber,A.L., Simmons,S.K., Moonshine,A.L., Ji,Z., Busby,M.A., Shi,X., Jacques,J., Lancaster,M.A., Pan,J.Q. *et al.* (2018) Comprehensive comparative analysis of 5′-end RNA-sequencing methods. *Nat. Methods*, **15**, 505–511.
28. Kawaji,H., Lizio,M., Itoh,M., Kanamori-Katayama,M., Kaiho,A., Nishiyori-Sueki,H., Shin,J.W., Kojima-Ishiyama,M., Kawano,M., Murata,M. *et al.* (2014) Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res.*, **24**, 708–717.

29. Pal,S., Gupta,R., Kim,H., Wickramasinghe,P., Baubet,V., Showe,L.C., Dahmane,N. and Davuluri,R.V. (2011) Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res.*, **21**, 1260–1272.
30. Valen,E., Pascarella,G., Chalk,A., Maeda,N., Kojima,M., Kawazu,C., Murata,M., Nishiyori,H., Lazarevic,D., Motti,D. *et al.* (2009) Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res.*, **19**, 255–265.
31. Arner,E., Daub,C.O., Vitting-Seerup,K., Andersson,R., Lilje,B., Drablos,F., Lennartsson,A., Rönnerblad,M., Hrydziuszko,O., Vitezic,M. *et al.* (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, **347**, 1010–1014.
32. Petersen,J. and Russell,P. (2016) Growth and the environment of schizosaccharomyces pombe. *Cold Spring Harb. Protoc.*, **2016**, doi:10.1101/pdb.top079764.
33. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
34. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
35. Hahne,F. and Ivanek,R. (2016) Visualizing genomic data using gviz and bioconductor. *Methods Mol. Biol.*, **1418**, 335–351.
36. Wagih,O. (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, **33**, 3645–3647.
37. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.
38. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
39. Persson,J., Steglich,B., Smialowska,A., Boyd,M., Bornholdt,J., Andersson,R., Schurra,C., Arcangioli,B., Sandelin,A., Nielsen,O. *et al.* (2016) Regulating retrotransposon activity through the use of alternative transcription start sites. *EMBO Rep.*, **17**, 753–768.
40. DeGennaro,C.M., Alver,B.H., Marguerat,S., Stepanova,E., Davis,C.P., Bähler,J., Park,P.J. and Winston,F. (2013) Spt6 regulates intragenic and antisense transcription, nucleosome positioning, and histone modifications genome-wide in fission yeast. *Mol. Cell. Biol.*, **33**, 4779–4792.
41. Booth,G.T., Wang,I.X., Cheung,V.G. and Lis,J.T. (2016) Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast. *Genome Res.*, **26**, 799–811.
42. Wery,M., Gautier,C., Descrimes,M., Yoda,M., Vennin-Rendos,H., Migeot,V., Gautheret,D., Hermand,D. and Morillon,A. (2017) Native elongating transcript sequencing reveals global anti-correlation between sense and antisense nascent transcription in fission yeast. *RNA*, **24**, 196–208.
43. Eser,P., Wachutka,L., Maier,K.C., Demel,C., Boroni,M., Iyer,S., Cramer,P. and Gagneur,J. (2016) Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. *Mol. Syst. Biol.*, **12**, 857–857.
44. Li,H., Hou,J., Bai,L., Hu,C., Tong,P., Kang,Y., Zhao,X. and Shao,Z. (2015) Genome-wide analysis of core promoter structures in Schizosaccharomyces pombe with DeepCAGE. *RNA Biol.*, **12**, 525–537.
45. Bornholdt,J., Saber,A.T., Lilje,B., Boyd,M., Jørgensen,M., Chen,Y., Vitezic,M., Jacobsen,N.R., Poulsen,S.S., Berthing,T. *et al.* (2017) Identification of gene transcription start sites and enhancers responding to pulmonary carbon nanotube exposure in vivo. *ACS Nano*, **11**, 3597–3613.
46. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2009) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Nat. Genet.*, **26**, 139–140.
47. Ritchie,M.E., Phipson,B., Wu,Di, Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47–e47.
48. Wilhelm,B.T., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C.J., Rogers,J. and Bähler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.
49. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-Regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
50. Jeffares,D.C., Rallis,C., Rieux,A., Speed,D., Převorovský,M., Mourier,T., Marsellach,F.X., Iqbal,Z., Lau,W., Cheng,T.M.K. *et al.* (2015) The genomic and phenotypic diversity of Schizosaccharomyces pombe. *Nat. Genet.*, **47**, 235–241.
51. Venables,W.N. and Ripley,B.D. (2002) *Modern Applied Statistics with S*. Springer New York, NY.
52. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2015) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
53. Boyd,M., Coskun,M., Lilje,B., Andersson,R., Hoof,I., Bornholdt,J., Dahlgaard,K., Olsen,J., Vitezic,M., Bjerrum,J.T. *et al.* (2014) Identification of TNF-α-responsive promoters and enhancers in the intestinal epithelial cell model Caco-2. *DNA Res.*, **21**, 569–583.
54. Smale,S.T. and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.
55. Ponjavic,J., Lenhard,B., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y. and Sandelin,A. (2006) Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol.*, **7**, R78.
56. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
57. Lantermann,A.B., Straub,T., Strålfors,A., Yuan,G.-C., Ekwall,K. and Korber,P. (2010) Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae. *Nat. Struct. Mol. Biol.*, **17**, 251–257.
58. Moyle-Heyrman,G., Zaichuk,T., Xi,L., Zhang,Q., Uhlenbeck,O.C., Holmgren,R., Widom,J. and Wang,J.-P. (2013) Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 20158–20163.
59. González,S., García,A., Vázquez,E., Serrano,R., Sánchez,M., Quintales,L. and Antequera,F. (2016) Nucleosomal signatures impose nucleosome positioning in coding and noncoding sequences in the genome. *Genome Res.*, **26**, 1532–1543.
60. Lenhard,B., Sandelin,A. and Carninci,P. (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.*, **13**, 233–245.
61. Jensen,T.H., Jacquier,A. and Libri,D. (2013) Dealing with pervasive transcription. *Mol. Cell*, **52**, 473–484.
62. Dutrow,N., Nix,D.A., Holt,D., Milash,B., Dalley,B., Westbroek,E., Parnell,T.J. and Cairns,B.R. (2008) Dynamic transcriptome of Schizosaccharomyces pombe shown by RNA-DNA hybrid mapping. *Nat. Genet.*, **40**, 977–986.
63. Shetty,A., Kallgren,S.P., Demel,C., Maier,K.C., Spatt,D., Alver,B.H., Cramer,P., Park,P.J. and Winston,F. (2017) Spt5 plays vital roles in the control of sense and antisense transcription elongation. *Mol. Cell*, **66**, 77–88.
64. Boyd,M., Thodberg,M., Vitezic,M., Bornholdt,J., Vitting-Seerup,K., Chen,Y., Coskun,M., Li,Y., Lo,B.Z.S., Klausen,P. *et al.* (2018) Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies. *Nat. Commun.*, **9**, 1661.
65. Sugimoto,A., Iino,Y., Maeda,T., Watanabe,Y. and Yamamoto,M. (1991) Schizosaccharomyces pombe ste11+encodes a transcription factor with an HMG motif that is a critical regulatorof sexual development. *Genes & Development*, **5**, 1990–1999.
66. Kjærulff,S., Lautrup-Larsen,I., Truelsen,S., Pedersen,M. and Nielsen,O. (2005) Constitutiveactivation of the fission yeast pheromone-responsive pathway induces ectopicmeiosis and reveals

ste11 as a mitogen-activated protein kinase target. *Mol Cell Biol.*, **25**, 2045–2059.
67. Leong,H.S., Dawson,K., Wirth,C., Li,Y., Connolly,Y., Smith,D.L., Wilkinson,C.R.M. and Miller,C.J. (2014) A global non-coding RNA system modulates fission yeast protein levels in response to stress. *Nat Commun.*, **5**, 3947.
68. Taricani,L., Feilotter,H.E., Weaver,C. and Young,P.G., , and (2001) Expression of hsp16 in response to nucleotide depletion is regulated via the spc1 MAPK pathway in Schizosaccharomyces pombe. *Nucleic Acids Res.*, **29**, 3030–3040.
69. Maeta,K., Nomura,W., Takatsume,Y., Izawa,S. and Inoue,Y. (2007) Green tea polyphenols function asprooxidants to activate oxidative-stress-responsive transcription factors inyeasts. *Appl. Environ. Microbiol.*, **73**, 572–580.
70. Marion,R.M., Regev,A., Segal,E., Barash,Y., Koller,D., Friedman,N. and O'Shea,E.K. (2004) Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 14315–14322.
71. Kim,L., Hoe,K.-L., Yu,Y.M., Yeon,J.-H. and Maeng,P.J. (2011) The fission yeast GATA factor, Gaf1, modulates sexual development via direct down-regulation of ste11+ expression in response to nitrogen starvation. *PLoS One*, **7**, e42409–e42409.
72. Kainou,T., Shinzato,T., Sasaki,K., Mitsui,Y., Giga-Hama,Y., Kumagai,H. and Uemura,H. (2006) Spsgt1, a new essential gene of Schizosaccharomyces pombe, is involved in carbohydrate metabolism. *Yeast*, **23**, 35–53.
73. Yazdi,P.G., Pedersen,B.A., Taylor,J.F., Khattab,O.S., Chen,Y.-H., Chen,Y., Jacobsen,S.E. and Wang,P.H. (2015) Increasing nucleosome occupancy is correlated with an increasing mutation rate so long as DNA repair machinery is intact. *PLoS One*, **10**, e0136574.
74. Davuluri,R.V., Suzuki,Y., Sugano,S., Plass,C. and Huang,T.H.M. (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet.*, **24**, 167–177.
75. Wu,Y.-H., Taggart,J., Song,P.X., MacDiarmid,C. and Eide,D.J. (2016) An MSC2 Promoter-lacZ fusion gene reveals Zinc-Responsive changes in sites of transcription initiation that occur across the yeast genome. *PLoS One*, **11**, e0163256.
76. Cheung,V., Chua,G., Batada,N.N., Landry,C.R., Michnick,S.W., Hughes,T.R. and Winston,F. (2008) Chromatin- and transcription-related factors repress transcription from within coding regions throughout the saccharomyces cerevisiae genome. *PLoS Biol.*, **6**, e277.
77. Zhou,S., Sternglanz,R. and Neiman,A.M. (2017) Developmentally regulated internal transcription initiation during meiosis in budding yeast. *PLoS One*, **12**, e0188001.
78. McKnight,K., Liu,H. and Wang,Y. (2014) Replicative stress induces intragenic transcription of the ASE1 gene that negatively regulates Ase1 activity. *Curr. Biol.*, **24**, 1101–1106.
79. Lee,J., Kwon,E.-S., Kim,D.-W., Cha,J. and Roe,J.-H. (2002) Regulation and the role of Cu,Zn-containing superoxide dismutase in cell cycle progression of Schizosaccharomyces pombe. *Biochem. Biophys. Res. Commun.*, **297**, 854–862.
80. Tanaka,K. and Russell,P. (2004) Cds1 phosphorylation by Rad3-Rad26 kinase is mediated by forkhead-associated domain interaction with Mrc1. *J. Biol. Chem.*, **279**, 32079–32086.
81. Vitting-Seerup,K. and Sandelin,A. (2017) The landscape of isoform switches in human cancers. *Mol. Cancer Res.*, **15**, 1206–1220.